RESEARCH CENTRE

Grenoble - Rhône-Alpes

2021

ACTIVITY REPORT

Project-Team

THOTH

Learning visual models from large-scale data

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

DOMAIN

Perception, Cognition and Interaction

THEME

Vision, perception and multimedia interpretation

Contents

Project-Team THOTH

Creation of the Project-Team: 2016 March 01

Keywords

Computer sciences and digital sciences

A3.4. – Machine learning and statistics

A5.3. – Image processing and analysis

A5.4. – Computer vision

A5.9. – Signal processing

A6.2.6. – Optimization

A8.2. – Optimization

A9.2. – Machine learning

A9.3. – Signal analysis

A9.7. – AI algorithmics

Other research topics and application domains

B9.5.6. – Data science

1 Team members, visitors, external collaborators

Research Scientists

- Julien Mairal [Team leader, Inria, Researcher, HDR]
- Karteek Alahari [Inria, Researcher, HDR]
- Pierre Gaillard [Inria, Researcher]

Faculty Member

- Jocelyn Chanussot [Institut polytechnique de Grenoble, Professor, HDR]

Post-Doctoral Fellows

- Michael Arbel [Inria, from Apr 2021]
- Heeseung Kwon [Inria, from Sep 2021]
- Huu Dien Khue Le [Inria, from Jun 2020]
- Romain Ménégaux [Univ. Grenoble Alpes, from Nov. 2021]
- Margot Selosse [Univ Grenoble Alpes, from Nov. 2020]

PhD Students

- Minttu Alakuijala [Google, CIFRE]
- Florent Bartoccioni [VALEO, CIFRE]
- Gaspard Beugnot [Inria, from Apr 2021]
- Jules Bourcier [Preligens SAS, from Jun 2021]
- Mathilde Caron [Facebook, CIFRE]
- Camila Fernandez Morales [Bell Labs (Alcatel)]
- Valentin Gabeur [Google, CIFRE]
- Ekaterina Iakovleva [Univ Grenoble Alpes]
- Zhiqi Kang [Inria, from Oct 2021]
- Roman Klokov [Inria, until May 2021]
- Bruno Lecouat [Inria]
- Hubert Leterme [Univ Grenoble Alpes]
- Lina Mezghani [Facebook, CIFRE]
- Gregoire Mialon [Inria, until Sep 2021]
- Alexander Pashevich [Google, until Jun 2021]
- Mert Bulent Sariyildiz [Naver Labs Europe, CIFRE]
- Vladyslav Sydorov [Inria, until Mar 2021]
- Houssam Zenati [Criteo]
- Alexandre Zouaoui [Inria]

Technical Staff

- Gaspard Beugnot [Inria, Engineer, until Mar 2021]
- Theo Bodrito [Inria, Engineer, until Oct 2021]
- Dexiong Chen [Inria, Engineer, until May 2021]
- Timothee Darcet [Inria, Engineer, from Oct 2021]
- Juliette Marrie [Univ Grenoble Alpes, Engineer, from Oct 2021]
- Gedeon Muhawenayo [Inria, Engineer, from Dec 2020]
- Thomas Ryckeboer [Inria, Engineer, from Oct 2021]

Interns and Apprentices

- Mohammed Almarakby [Criteo, from Mar 2021 until Aug 2021]
- Pierre Andre Crepon [Ecole normale supérieure Paris-Saclay, from Apr 2021 until Aug 2021]
- Aymeric Darolles [Sorbonne Université, from May 2021 until Oct 2021]
- Jerome Taupin [Ecole normale supérieure Paris-Saclay, from Jun 2021 until Jul 2021]

Administrative Assistant

- Nathalie Gillot [Inria]

Visiting Scientists

- Pia Bideau [Université technique de Berlin - Allemagne, from Sep 2021]
- Enrico Fini [Université de Trente - Italie, from Sep 2021]
- David Alejandro Jimenez Sierra [Pontificia Universidad Javeriana Cali, Columbia, from Nov 2021]
- Huan Ni [School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China, until Jan 2021]
- Jie Xie [College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China, from Jan 2021]
- Yuxuan Zheng [Université normale de la Chine de l'Est (ECNU) Shanghai, from Apr 2021]

2 Overall objectives

Thoth is a computer vision and machine learning team. Our initial goal was to develop machine learning models for analyzing the massive amounts of visual data that are currently available on the web. Then, the focus of the team has become more diverse. More precisely, we share a common objective of developing machine learning models that are robust and efficient (in terms of computational cost and data requirements).

Our main research directions are the following ones:

- **visual understanding from limited annotations and data:** Many state-of-the-art computer vision models are typically trained on a huge corpus of fully annotated data. We want to reduce the cost by developing new algorithms for unsupervised, self-supervised, continual, or incremental learning.

- **efficient deep learning models, from theory to applications**: We want to invent a new generation of machine learning models (in particular deep learning) with theoretical guarantees, efficient algorithms, and a wide range of applications. We develop for instance models for images, videos, graphs, or sequences.

- **statistical machine learning and optimization:** we are also developing efficient machine learning methods, with a focus on stochastic optimization for processing large-scale data, and online learning.

- **pluri-disciplinary collaborations:** Machine learning being at the crossing of several disciplines, we have successfully conducted collaborations in scientific domains that are relatively far from our domains of expertise. These fields are producing massive amounts of data and are in dire needs of efficient tools to make predictions or interpretations. For example, we have had the chance to collaborate with many colleagues from natural language processing, robotics, neuroimaging, computational biology, genomics, astrophysics for exoplanet detections, and we are currently involved in several remote sensing and hyperspectral imaging projects thanks to Jocelyn Chanussot (hosted by Thoth from 2019 to 2022).

3 Research program

3.1 Designing and learning structured models

The task of understanding image and video content has been interpreted in several ways over the past few decades, namely image classification, detecting objects in a scene, recognizing objects and their spatial extents in an image, recovering scene geometry. However, addressing all these problems individually provides us with a partial understanding of the scene at best, leaving much of the visual data unexplained.

One of the main goals of this research axis is to go beyond the initial attempts that consider only a subset of tasks jointly, by developing novel models for a more complete understanding of scenes to address all the component tasks. We propose to incorporate the structure in image and video data explicitly into the models. In other words, our models aim to satisfy the complex sets of constraints that exist in natural images and videos. Examples of such constraints include: (i) relations between objects, like signs for shops indicate the presence of buildings, (ii) higher-level semantic relations involving the type of scene, geographic location, and the plausible actions as a global constraint, e.g., an image taken at a swimming pool is unlikely to contain cars, (iii) relating objects occluded in some of the video frames to content in other frames, where they are more clearly visible as the camera or the object itself move, with the use of long-term trajectories and video object proposals.

This research axis will focus on two topics. The first is developing deep features for video. This involves designing rich features available in the form of long-range temporal interactions among pixels in a video sequence to learn a representation that is truly spatio-temporal in nature. The second topic is aimed at learning models that capture the relationships among several objects and regions in a single image scene, and additionally, among scenes in the case of an image collection or a video. The main scientific challenges in this topic stem from learning the structure of the probabilistic graphical model as well as the parameters of the cost functions quantifying the relationships among its entities. In the following we will present work related to all these three topics and then elaborate on our research directions.

- **Deep features for vision.** Deep learning models provide a rich representation of complex objects but in return have a large number of parameters. Thus, to work well on difficult tasks, a large amount of data is required. In this context, video presents several advantages: objects are observed from a large range of viewpoints, motion information allows the extraction of moving objects and parts, and objects can be differentiated by their motion patterns. We initially plan to develop deep features for videos that incorporate temporal information at multiple scales. We then plan to further exploit the rich content in video by incorporating additional cues such as minimal prior knowledge of the object of interest, with the goal of learning a representation that is more appropriate for video understanding. In other words, a representation that is learned from video data and targeted at specific applications.

- **Structured models.** The interactions among various elements in a scene, such as the objects and regions in it, the motion of object parts or entire objects themselves, form a key element for understanding image or video content. These rich cues define the structure of visual data and how it evolves spatio-temporally. We plan to develop a novel graphical model to exploit this structure. The main components in this graphical model are spatio-temporal regions (in the case of video or simply image regions), which can represent object parts or entire objects themselves, and the interactions among several entities. The dependencies among the scene entities are defined with a higher order or a global cost function. A higher order constraint is a generalization of the pairwise interaction term, and is a cost function involving more than two components in the scene, e.g., several regions, whereas a global constraint imposes a cost term over the entire image or vide such as a prior knowledge on the number of people expected in the scene. The constraints we plan to include generalize several existing methods, which are limited to pairwise interactions or a small restrictive set of higher-order costs. In addition to learning the parameters of these novel functions, we will focus on learning the structure of the graph itself—a challenging problem that is seldom addressed in current approaches. This provides an elegant way to go beyond state-of-the-art deep learning methods, which are limited to learning the high-level interaction among parts of an object, by learning the relationships among objects.

3.2 Learning of visual models from minimal supervision

Today's approaches to visual recognition learn models for a limited and fixed set of visual categories with fully supervised classification techniques. This paradigm has been adopted in the early 2000's, and within it enormous progress has been made over the last decade.

The scale and diversity in today's large and growing image and video collections (such as, e.g., broadcast archives, and personal image/video collections) call for a departure from the current paradigm. This is the case because to answer queries about such data, it is unfeasible to learn the models of visual content by manually and precisely annotating every relevant concept, object, scene, or action category in a representative sample of everyday conditions. For one, it will be difficult, or even impossible to decide a-priori what are the relevant categories and the proper granularity level. Moreover, the cost of such annotations would be prohibitive in most application scenarios. One of the main goals of the Thoth project-team is to develop a new framework for learning visual recognition models by actively exploring large digital image and video sources (off-line archives as well as growing on-line content), and exploiting the weak supervisory signal provided by the accompanying metadata (such as captions, keywords, tags, subtitles, or scripts) and audio signal (from which we can for example extract speech transcripts, or exploit speaker recognition models).

Textual metadata has traditionally been used to index and search for visual content. The information in metadata is, however, typically sparse (e.g., the location and overall topic of newscasts in a video archive [1]) and noisy (e.g., a movie script may tell us that two persons kiss in some scene, but not when, and the kiss may occur off the screen or not have survived the final cut). For this reason, metadata search should be complemented by visual content based search, where visual recognition models are used to localize content of interest that is not mentioned in the metadata, to increase the usability and value of image/video archives. *The key insight that we build on in this research axis is that while the metadata for a single image or video is too sparse and noisy to rely on for search, the metadata associated with large video and image databases collectively provide an extremely versatile source of information to learn visual recognition models.* This form of "embedded annotation" is rich, diverse and abundantly available. Mining these correspondences from the web, TV and film archives, and online consumer generated content sites such as Flickr, Facebook, or YouTube, guarantees that the learned models are representative for many different situations, unlike models learned from manually collected fully supervised training data sets which are often biased.

The approach we propose to address the limitations of the fully supervised learning paradigm aligns with "Big Data" approaches developed in other areas: we rely on the orders-of-magnitude-larger training sets that have recently become available with metadata to compensate for less explicit forms of supervision. This will form a sustainable approach to learn visual recognition models for a much larger set

[1] For example at the Dutch national broadcast archive Netherlands Institute of Sound and Vision, with whom we collaborated in the EU FP7 project AXES, typically one or two sentences are used in the metadata to describe a one hour long TV program.

of categories with little or no manual intervention. Reducing and ultimately removing the dependency on manual annotations will dramatically reduce the cost of learning visual recognition models. This in turn will allow such models to be used in many more applications, and enable new applications based on visual recognition beyond a fixed set of categories, such as natural language based querying for visual content. This is an ambitious goal, given the sheer volume and intrinsic variability of the every day visual content available on-line, and the lack of a universally accepted formalism for modeling it. Yet, the potential payoff is a breakthrough in visual object recognition and scene understanding capabilities.

This research axis is organized into the following three sub-tasks:

- **Weakly supervised learning.** For object localization we will go beyond current methods that learn one category model at a time and develop methods that learn models for different categories concurrently. This allows "explaining away" effects to be leveraged, i.e., if a certain region in an image has been identified as an instance of one category, it cannot be an instance of another category at the same time. For weakly supervised detection in video we will consider detection proposal methods. While these are effective for still images, recent approaches for the spatio-temporal domain need further improvements to be similarly effective. Furthermore, we will exploit appearance and motion information jointly over a set of videos. In the video domain we will also continue to work on learning recognition models from subtitle and script information. The basis of leveraging the script data which does not have a temporal alignment with the video is to use matches in the narrative in the script and the subtitles (which do have a temporal alignment with the video). We will go beyond simple correspondences between names and verbs relating to self-motion, and match more complex sentences related to interaction with objects and other people. To deal with the limited number of occurrences of such actions in a single movie, we will consider approaches that learn action models across a collection of movies.

- **Online learning of visual models.** As a larger number of visual category models is being learned, online learning methods become important, since new training data and categories will arrive over time. We will develop online learning methods that can incorporate new examples for existing category models, and learn new category models from few examples by leveraging similarity to related categories using multi-task learning methods. Here we will develop new distance-based classifiers and attribute and label embedding techniques, and explore the use of NLP techniques such as skipgram models to automatically determine between which classes transfer should occur. Moreover, NLP will be useful in the context of learning models for many categories to identify synonyms, and to determine cases of polysemy (e.g. jaguar car brand v.s. jaguar animal), and merge or refine categories accordingly. Ultimately this will result in methods that are able to learn an"encyclopedia" of visual models.

- **Visual search from unstructured textual queries.** We will build on recent approaches that learn recognition models on-the-fly (as the query is issued) from generic image search engines such as Google Images. While it is feasible to learn models in this manner in a matter of seconds, it is challenging to use the model to retrieve relevant content in real-time from large video archives of more than a few thousand hours. To achieve this requires feature compression techniques to store visual representations in memory, and cascaded search techniques to avoid exhaustive search. This approach, however, leaves untouched the core problem of how to associate visual material with the textual query in the first place. The second approach we will explore is based on image annotation models. In particular we will go beyond image-text retrieval methods by using recurrent neural networks such as Elman networks or long short-term memory (LSTM) networks to generate natural language sentences to describe images.

3.3 Large-scale learning and optimization

We have entered an era of massive data acquisition, leading to the revival of an old scientific utopia: it should be possible to better understand the world by automatically converting data into knowledge. It is also leading to a new economic paradigm, where data is a valuable asset and a source of activity. Therefore, developing scalable technology to make sense of massive data has become a strategic issue. Computer vision has already started to adapt to these changes.

In particular, very high-dimensional models such as deep networks are becoming highly popular and successful for visual recognition. This change is closely related to the advent of big data. On the one hand, these models involve a huge number of parameters and are rich enough to represent well complex objects such as natural images or text corpora. On the other hand, they are prone to overfitting (fitting too closely to training data without being able to generalize to new unseen data) despite regularization; to work well on difficult tasks, they require a large amount of labeled data that has been available only recently. Other cues may explain their success: the deep learning community has made significant engineering efforts, making it possible to learn in a day on a GPU large models that would have required weeks of computations on a traditional CPU, and it has accumulated enough empirical experience to find good hyper-parameters for its networks.

To learn the huge number of parameters of deep hierarchical models requires scalable optimization techniques and large amounts of data to prevent overfitting. This immediately raises two major challenges: how to learn without large amounts of labeled data, or with weakly supervised annotations? How to efficiently learn such huge-dimensional models? To answer the above challenges, we will concentrate on the design and theoretical justifications of deep architectures including our recently proposed deep kernel machines, with a focus on weakly supervised and unsupervised learning, and develop continuous and discrete optimization techniques that push the state of the art in terms of speed and scalability.

This research axis will be developed into three sub-tasks:

- **Deep kernel machines for structured data.** Deep kernel machines combine advantages of kernel methods and deep learning. Both approaches rely on high-dimensional models. Kernels implicitly operate in a space of possibly infinite dimension, whereas deep networks explicitly construct high-dimensional nonlinear data representations. Yet, these approaches are complementary: Kernels can be built with deep learning principles such as hierarchies and convolutions, and approximated by multilayer neural networks. Furthermore, kernels work with structured data and have well understood theoretical principles. Thus, a goal of the Thoth project-team is to design and optimize the training of such deep kernel machines.

- **Large-scale parallel optimization.** Deep kernel machines produce nonlinear representations of input data points. After encoding these data points, a learning task is often formulated as a *large-scale convex optimization problem*; for example, this is the case for linear support vector machines, logistic regression classifiers, or more generally many empirical risk minimization formulations. We intend to pursue recent efforts for making convex optimization techniques that are dedicated to machine learning more scalable. Most existing approaches address scalability issues either in model size (meaning that the function to minimize is defined on a domain of very high dimension), or in the amount of training data (typically, the objective is a large sum of elementary functions). There is thus a large room for improvements for techniques that jointly take these two criteria into account.

- **Large-scale graphical models.** To represent structured data, we will also investigate graphical models and their optimization. The challenge here is two-fold: designing an adequate cost function and minimizing it. While several cost functions are possible, their utility will be largely determined by the efficiency and the effectiveness of the optimization algorithms for solving them. It is a combinatorial optimization problem involving billions of variables and is NP-hard in general, requiring us to go beyond the classical approximate inference techniques. The main challenges in minimizing cost functions stem from the large number of variables to be inferred, the inherent structure of the graph induced by the interaction terms (e.g., pairwise terms), and the high-arity terms which constrain multiple entities in a graph.

4 Application domains

4.1 Visual applications

Any solution to automatically understanding images and videos on a semantic level will have an immediate impact on a wide range of applications. For example:

- Semantic-level image and video access is highly relevant for visual search on the Web, in professional archives and personal collections.
- Visual data organization is applicable to organizing family photo and video albums as well as to large-scale information retrieval.
- Visual object recognition has potential applications ranging from autonomous driving, to service robotics for assistance in day-to-day activities as well as the medical domain.
- Real-time scene understanding is relevant for human interaction through devices such as HoloLens, Oculus Rift.

4.2 Pluri-disciplinary research

Machine learning is intrinsically pluri-disciplinary. By developing large-scale machine learning models and algorithms for processing data, the Thoth team became naturally involved in pluri-disciplinary collaborations that go beyond visual modelling. During the last few years, Thoth has conducted several collaborations in other fields such as neuroimaging, bioinformatics, natural language processing, and remote sensing.

5 Highlights of the year

5.1 Awards

- Maha Elbayad, former PhD student of the team, has received the EAMT (European Association for Machine Translation) best PhD award.
- Pierre Gaillard has received (with his co-authors) an outstanding paper award at NeurIPS 2021.
- Jocelyn Chanussot is Highly Cited Researcher (HCR - Thomson Reuters / Clarivate Analytics).
- Jocelyn Chanussot was appointed ELLIS fellow and fellow of the Asia-Pacific Artificial Intelligence Association (AAIA).

6 New software and platforms

You will find below the new software packages developed by the Thoth team.

6.1 New software

6.1.1 Cyanure

Name: Cyanure: An Open-Source Toolbox for Empirical Risk Minimization

Keyword: Machine learning

Functional Description: Cyanure is an open-source C++ software package with a Python interface. The goal of Arsenic is to provide state-of-the-art solvers for learning linear models, based on stochastic variance-reduced stochastic optimization with acceleration mechanisms and Quasi-Newton principles. Arsenic can handle a large variety of loss functions (logistic, square, squared hinge, multinomial logistic) and regularization functions (l2, l1, elastic-net, fused Lasso, multi-task group Lasso). It provides a simple Python API, which is very close to that of scikit-learn, which should be extended to other languages such as R or Matlab in a near future.

Release Contributions: version initiale

URL: http://thoth.inrialpes.fr/people/mairal/arsenic/welcome.html

Contact: Julien Mairal

Participant: Julien Mairal

6.1.2 OTK

Name: Optimal Transport Kernel Embedding

Keyword: Machine learning

Functional Description: Source code reproducing the results of the following paper

Grégoire Mialon*, Dexiong Chen*, Alexandre d'Aspremont, Julien Mairal. A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention. ICLR 2021.

URL: https://github.com/claying/OTK/

Contact: Julien Mairal

6.1.3 T3SC

Name: A Trainable Spectral-Spatial Sparse Coding Model for Hyperspectral Image Restoration

Keyword: Machine learning

Functional Description: This is the source code for reproducing the results of the paper

A Trainable Spectral-Spatial Sparse Coding Model for Hyperspectral Image Restoration (Neurips 2021)

URL: https://github.com/inria-thoth/T3SC/

Contact: Théo Bodrito

6.1.4 GraphIT

Name: GraphiT: Encoding Graph Structure in Transformers

Keyword: Machine learning

Functional Description: This is the source code for reproducing the results of the paper

Grégoire Mialon*, Dexiong Chen*, Margot Selosse*, Julien Mairal. GraphiT: Encoding Graph Structure in Transformers.

URL: https://github.com/inria-thoth/GraphiT/

Contact: Grégoire Mialon

6.1.5 cog-eval

Name: The ImageNet-CoG Benchmark

Keywords: Computer vision, Evaluation

Functional Description: Code repository for the ImageNet-CoG Benchmark introduced in the paper "Concept Generalization in Visual Representation Learning" (ICCV 2021). It contains code for reproducing all the experiments reported in the paper, as well as instructions on how to evaluate any custom model on the ImageNet-CoG Benchmark.

URL: https://github.com/naver/cog

Contact: Karteek Alahari

6.1.6 Multi-modal transformer

Name: Multi-modal Transformer for Video Retrieval

Keywords: Computer vision, Video retrieval, Transformer

Functional Description: The task of retrieving video content relevant to natural language queries plays a critical role in effectively handling internet-scale datasets. Most of the existing methods for this caption-to-video retrieval problem do not fully exploit cross-modal cues present in video. Furthermore, they aggregate per-frame visual features with limited or no temporal information. In this paper, we present a multi-modal transformer to jointly encode the different modalities in video, which allows each of them to attend to the others. The transformer architecture is also leveraged to encode and model the temporal information. On the natural language side, we investigate the best practices to jointly optimize the language embedding together with the multi-modal transformer. This is an implementation of our novel framework, which was presented at ECCV 2020.

URL: http://thoth.inrialpes.fr/research/mmt/

Contact: Karteek Alahari

Partner: Google

7 New results

7.1 Visual Recognition

Emerging Properties in Self-Supervised Vision Transformers

Participants: Mathilde Caron, Hugo Touvron, Ishan Mishra, Hervé Jégou, Piotr Bojanowski, Julien Mairal.

In [6], we question if self-supervised learning provides new properties to Vision Transformer (ViT) that stand out compared to convolutional networks (convnets). Beyond the fact that adapting self-supervised methods to this architecture works particularly well, we make the following observations: first, self-supervised ViT features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets. Second, these features are also excellent k-NN classifiers, reaching 78.3% top-1 on ImageNet with a small ViT. Our study also underlines the importance of momentum encoder, multi-crop training, and the use of small patches with ViTs. We implement our findings into a simple self-supervised method, called DINO, which we interpret as a form of self-distillation with no labels. We show the synergy between DINO and ViTs by achieving 80.1% top-1 on ImageNet in linear evaluation with ViT-Base.

Concept generalization in visual representation learning

Participants: Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, Karteek Alahari.

In this work [20], we study concept generalization, i.e., the extent to which models trained on a set of (seen) visual concepts can be used to recognize a new set of (unseen) concepts. Although this paradigm is a popular way of evaluating visual representations, the choice of which unseen concepts to use is usually made arbitrarily, and independently from the seen concepts used to train representations, thus ignoring any semantic relationships between the two. We argue that semantic relationships between seen and unseen concepts affect generalization performance and propose ImageNet-CoG (see Figure 2), a novel benchmark on the ImageNet dataset that enables measuring concept generalization in a principled way. Our benchmark leverages expert knowledge that comes from WordNet in order to define a sequence

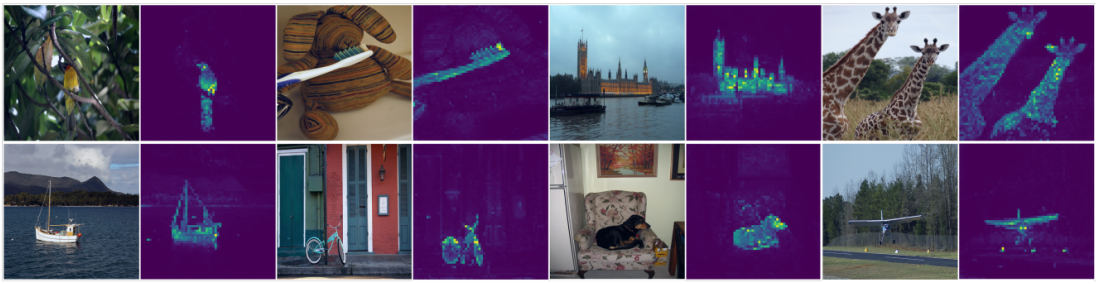

Figure 1: Self-attention from a Vision Transformer with 8 × 8 patches trained with no supervision. We look at the self-attention of the [CLS] token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

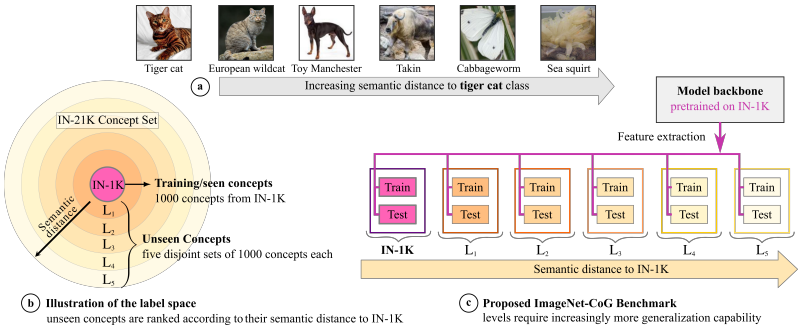

Figure 2: **An overview of our ImageNet Concept Generalization (ImageNet-CoG) benchmark.** (a) An example of five concepts from the ImageNet-21K dataset (IN-21K), ranked by increasing *semantic* distance (decreasing Lin similarity) to the ImageNet-1K (IN-1K) dataset concept "Tiger cat". (b) We rank the 21K concepts of IN-21K according to their semantic distance to the 1000 concepts of IN-1K and split the ranked list to extract 5 groups of 1000 concepts. We refer to the five IN-1K-sized datasets of increasing semantic distance from IN-1K as *concept generalization levels*, denoted as $L_{1/2/3/4/5}$. (c) The proposed ImageNet-CoG benchmark uses a model trained on IN-1K as a feature extractor and evaluates its concept generalization capabilities by learning linear classifiers for each level of more and more challenging unseen concepts.

of unseen ImageNet concept sets that are semantically more and more distant from the ImageNet-1K subset, a ubiquitous training set. Under the prism of concept generalization, we analyse a variety of supervised, semi-supervised and self-supervised models with different backbone architectures, such as Convolutional Neural Networks, Vision Transformers or architectures found by Neural Architecture Search models. Our large-scale study shows that our benchmark is able to uncover a number of interesting insights.

Regularized Frank-Wolfe for Dense CRFs: Generalizing Mean Field and Beyond

Participants: Đ. Khuê Lê-Huu, Karteek Alahari.

In [11], we introduce *regularized Frank-Wolfe*, a general and effective algorithm for inference and learning of dense conditional random fields (CRFs). The algorithm optimizes a nonconvex continuous relaxation of the CRF inference problem using vanilla Frank-Wolfe with approximate updates, which are equivalent to minimizing a regularized energy function. Our proposed method is a generalization of

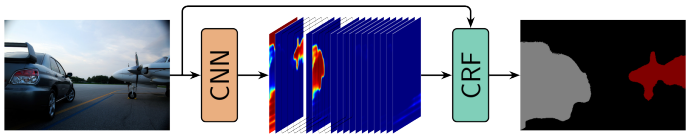

Figure 3: CRF as an end-to-end trainable component in a neural network.

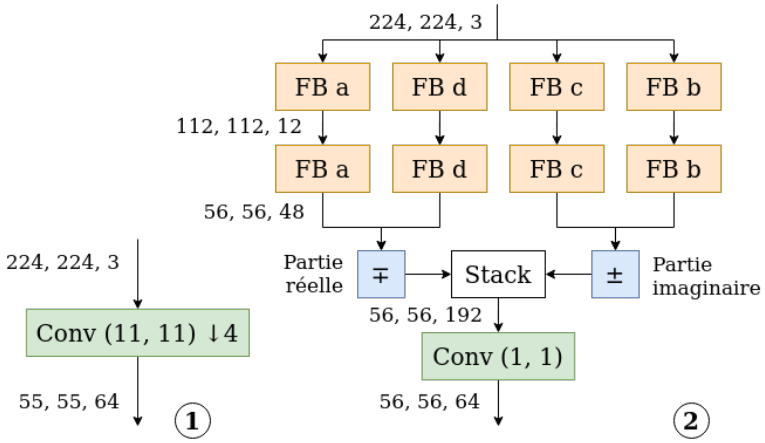

Figure 4: 1) First layer of AlexNet; 2) Proposed model, replacing the standard convolution layer. Each orange module ("FB", for *filter bank*) refers to one stage of discrete wavelet packet decomposition. Only the green modules ("Conv") contain trainable parameters. The numbers between each layer indicate the height and width of images as well as the number of channels.

existing algorithms such as mean field or concave-convex procedure. This perspective not only offers a unified analysis of these algorithms, but also allows an easy way of exploring different variants that potentially yield better performance. We illustrate this in our empirical results on standard semantic segmentation datasets, where several instantiations of our regularized Frank-Wolfe outperform mean field inference, both as a standalone component and as an end-to-end trainable layer in a neural network (Figure 3). We also show that dense CRFs, coupled with our new algorithms, produce significant improvements over strong CNN baselines.

Modélisation Parcimonieuse de CNNs avec des Paquets d'Ondelettes Dual-Tree

Participants: Hubert Leterme, Kévin Polisano, Valérie Perrier, Karteek Alahari.

We propose in [13] to improve the mathematical interpretability of convolutional neural networks (CNNs) for image classification. In this purpose, we replace the first layers of existing models such as AlexNet or ResNet by an operator containing the dual-tree wavelet packet transform, i.e., a redundant decomposition using complex and oriented waveforms. Figure 4 provides a schematic representation of one such architecture. Our experiments show that these modified networks behave very similarly to the original models once trained. The goal is then to study this operator from a theoretical point of view and to identify potential optimizations. We want to analyze its main properties such as directional selectivity, stability with respect to small shifts and rotations, thus retaining discriminant information while decreasing intra-class variability. This work is a step toward a more complete description of CNNs using well-defined mathematical operators, characterized by a small number of arbitrary parameters, making them easier to interpret.

LiDARTouch: Monocular metric depth estimation with a few-beam LiDAR

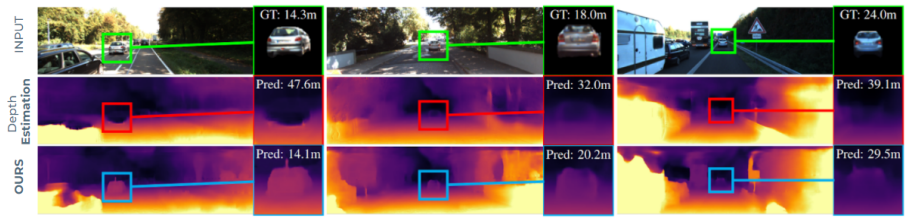

Figure 5: Mitigation of the infinite-depth problem.

Participants: Florent Bartoccioni, Eloi Zablocki, Patrick Pérez, Matthieu Cord, Karteek Alahari.

In this paper [23], we address the task of monocular depth prediction, a key component of many autonomous systems, by self-supervised deep learning. Existing methods are either fully-supervised with an additional expensive LiDAR (32 or 64 beams) as input or self-supervised with camera-only methods, much cheaper, but suffering from scale ambiguity and infinite depth problems. In contrast, we introduce LiDARTouch, a novel method combining a monocular camera with a cheap minimal 4-beam LiDAR input, typical of laser scanners currently used in the automotive industry. We introduce a new self-supervision scheme to leverage this very sparse LiDAR input at three complementary levels. While being extremely sparse, we show that the use of a few-beam LiDAR alleviate the scaling ambiguity and infinite depth problems that camera-only methods suffer from. We also reach competitive performances with respect to fully-supervised depth completion methods while being significantly cheaper and more annotation friendly. Our method can be trained on any domain with no modification, and it can thus bring accurate and metric depth estimation at a vehicle fleet scale. In Figure 5, we present three examples along with selected close-ups highlighting the infinite-depth problem. For example, on the leftmost column, we observe a typical 'hole' in the depth map where previous 'Depth Estimation' method estimates a vehicle three times as far as its true distance. In contrast, by leveraging small touches of LIDAR we disambiguate the prediction and can accurately and safely handle moving objects with no relative motion, typical of cars in fluid traffic.

Masking Modalities for Cross-modal Video Retrieval

Participants: Valentin Gabeur, Arsha Nagrani, Chen Sun, Kateek Alahari, Cordelia Schmid.

Pre-training on large scale unlabelled datasets has shown impressive performance improvements in the fields of computer vision and natural language processing. Given the advent of large-scale instructional video datasets, a common strategy for pre-training video encoders is to use the accompanying speech as weak supervision. However, as speech is used to supervise the pre-training, it is never seen by the video encoder, which does not learn to process that modality. We address this drawback of current pre-training methods, which fail to exploit the rich cues in spoken language, in [21]. Our proposal is to pre-train a video encoder using all the available video modalities as supervision, namely, appearance, sound, and transcribed speech. We mask an entire modality in the input and predict it using the other two modalities. This encourages each modality to collaborate with the others, and our video encoder learns to process appearance and audio as well as speech 6. We show the superior performance of our "modality masking" pre-training approach for video retrieval on the How2R, YouCook2 and Condensed Movies datasets.

7.2 Statistical Machine Learning

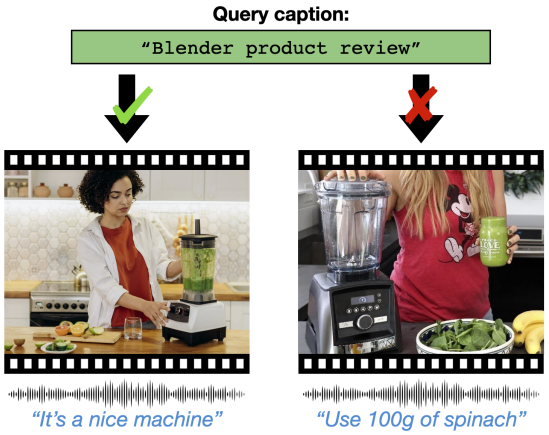

Figure 6: Speech is part of the story! Video retrieval methods that focus on visual inputs alone are likely to miss out on key information (e.g., while both the examples above contain a blender, the speech (in blue) helps identify the one for a product review). In this work, we focus on learning a video encoder to effectively process RGB and audio features, as well as transcribed speech from instructional videos online, through a novel modality masking method. Our approach learns from unlabelled videos, without the need for expensive manual captions.

A Contextual Bandit Bake-off

Participants: Alberto Bietti, Alekh Agarwal *(Microsoft Research)*, John Langford *(Microsoft Research)*.

Contextual bandit algorithms are essential for solving many real-world interactive machine learning problems. Despite multiple recent successes on statistically and computationally efficient methods, the practical behavior of these algorithms is still poorly understood. In [1], we leverage the availability of large numbers of supervised learning datasets to compare and empirically optimize contextual bandit algorithms, focusing on practical methods that learn by relying on optimization oracles from supervised learning. We find that a recent method using optimism under uncertainty works the best overall. A surprisingly close second is a simple greedy baseline that only explores implicitly through the diversity of contexts, followed by a variant of Online Cover which tends to be more conservative but robust to problem specification by design. Along the way, we also evaluate and improve several internal components of contextual bandit algorithm design. Overall, this is a thorough study and review of contextual bandit methodology.

Beyond Tikhonov: Faster Learning with Self-Concordant Losses via Iterative Regularization

Participants: Gaspard Beugnot, Julien Mairal, Alessandro Rudi.

The theory of spectral filtering is a remarkable tool to understand the statistical properties of learning with kernels. For least squares, it allows to derive various regularization schemes that yield faster convergence rates of the excess risk than with Tikhonov regularization. This is typically achieved by leveraging classical assumptions called source and capacity conditions, which characterize the difficulty of the learning task. In order to understand estimators derived from other loss functions, Marteau-Ferey et al. have extended the theory of Tikhonov regularization to generalized self concordant loss functions (GSC), which contain, *e.g.*, the logistic loss. In this paper [4], we go a step further and show that fast and optimal rates can be achieved for GSC by using the iterated Tikhonov regularization scheme, which is intrinsically related to the proximal point method in optimization, and overcomes the limitation of the classical Tikhonov regularization.

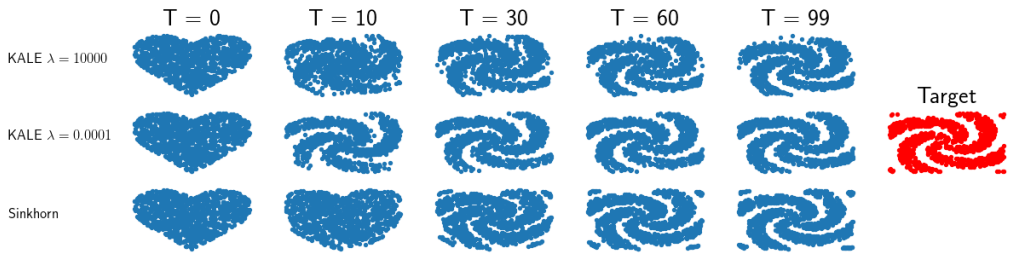

Figure 7: Shape Transfer using KALE flow.

KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support

Participants: Pierre Glaser, Michael Arbel, Arthur Gretton.

In [9], we study the gradient flow for a relaxed approximation to the Kullback-Leibler (KL) divergence between a moving source and a fixed target distribution. This approximation, termed the KALE (KL approximate lower-bound estimator), solves a regularized version of the Fenchel dual problem defining the KL over a restricted class of functions. When using a Reproducing Kernel Hilbert Space (RKHS) to define the function class, we show that the KALE continuously interpolates between the KL and the Maximum Mean Discrepancy (MMD). Like the MMD and other Integral Probability Metrics, the KALE remains well defined for mutually singular distributions. Nonetheless, the KALE inherits from the limiting KL a greater sensitivity to mismatch in the support of the distributions, compared with the MMD. These two properties make the KALE gradient flow particularly well suited when the target distribution is supported on a low-dimensional manifold. Under an assumption of sufficient smoothness of the trajectories, we show the global convergence of the KALE flow. We propose a particle implementation of the flow given initial samples from the source and the target distribution, which we use to empirically confirm the KALE's properties, see Figure 7.

Amortized implicit differentiation for stochastic bilevel optimization

Participants: Michael Arbel, Julien Mairal.

In [3], we study a class of algorithms for solving bilevel optimization problems in both stochastic and deterministic settings when the inner-level objective is strongly convex. Specifically, we consider algorithms based on inexact implicit differentiation and we exploit a warm-start strategy to amortize the estimation of the exact gradient, see Figure 8. We then introduce a unified theoretical framework inspired by the study of singularly perturbed systems (Habets, 1974) to analyze such amortized algorithms. By using this framework, our analysis shows these algorithms to match the computational complexity of oracle methods that have access to an unbiased estimate of the gradient, thus outperforming many existing results for bilevel optimization. We illustrate these findings on synthetic experiments and demonstrate the efficiency of these algorithms on hyper-parameter optimization experiments involving several thousands of variables.

Counterfactual Learning of Stochastic Policies with Continuous Actions

Participants: Houssam Zenati, Alberto Bietti, Matthieu Martin, Eustache Diemert, Julien Mairal.

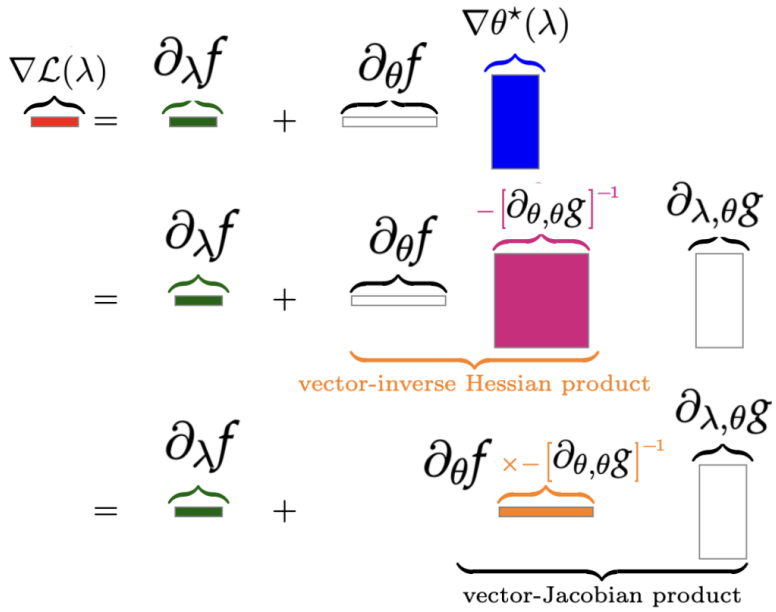

Figure 8: Implicit differentiation for computing the gradient of a bilevel objective.

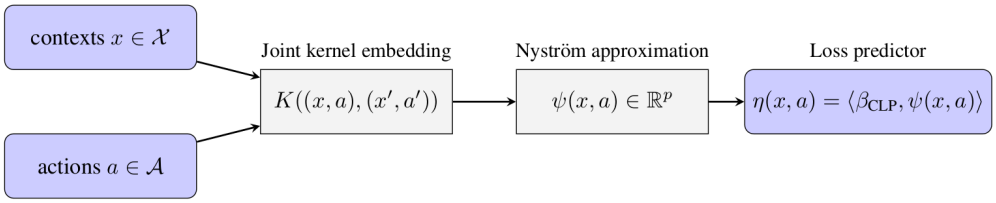

Figure 9: Illustration of the joint kernel embedding for the counterfactual loss predictor (CLP) and loss estimator.

Counterfactual reasoning from logged data has become increasingly important for many applications such as web advertising or healthcare. In this paper [26], we address the problem of counterfactual learning of stochastic policies with continuous actions, which raises difficult challenges about (i) data modelization, (ii) optimization, and (iii) evaluation on real data.

First, we introduce a modeling strategy based on a joint kernel embedding of contexts and actions, which overcomes the shortcomings of previous discretization strategies as shown in 9. Second, we empirically show that the optimization aspect of counterfactual learning is more important than previously thought, and we demonstrate the benefits of proximal point algorithms and differentiable estimators. Finally, we propose an evaluation protocol for offline policies in real-world logged systems, which is challenging since policies cannot be replayed on test data, and we release a new large-scale dataset along with multiple synthetic, yet realistic, evaluation setups.

A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip

Participants: Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Pierre Gaillard, Hadrien Hendrikx, Laurent Massoulié, Adrien Taylor.

In [7], we introduce the continuized Nesterov acceleration, a close variant of Nesterov acceleration whose variables are indexed by a continuous time parameter. The two variables continuously mix

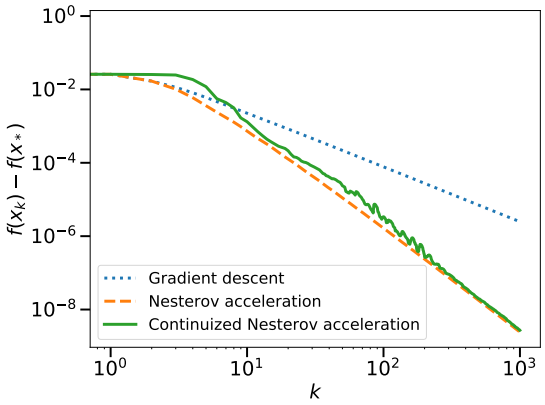

Figure 10: Illustration of the convergence of Gradient Descent, Continuized Nesterov acceleration and Nesterov acceleration on a convex objective.

following a linear ordinary differential equation and take gradient steps at random times. This continuized variant benefits from the best of the continuous and the discrete frameworks: as a continuous process, one can use differential calculus to analyze convergence and obtain analytical expressions for the parameters; and a discretization of the continuized process can be computed exactly with convergence rates similar to those of Nesterov original acceleration. We show that the discretization has the same structure as Nesterov acceleration, but with random parameters. We provide continuized Nesterov acceleration under deterministic as well as stochastic gradients, with either additive or multiplicative noise. Finally, using our continuized framework and expressing the gossip averaging problem as the stochastic minimization of a certain energy function, we provide the first rigorous acceleration of asynchronous gossip algorithms.

Mixability made efficient: Fast online multiclass logistic regression

Participants: Rémi Jézéquel, Pierre Gaillard, Alessandro Rudi.

Mixability has been shown to be a powerful tool to obtain algorithms with optimal regret. However, the resulting methods often suffer from high computational complexity which has reduced their practical applicability. For example, in the case of multiclass logistic regression, the aggregating forecaster (Foster et al., 2018) achieves a regret of $O(\log(Bn))$ whereas Online Newton Step achieves $O(e^B \log(n))$ obtaining a double exponential gain in B (a bound on the norm of comparative functions). However, this high statistical performance is at the price of a prohibitive computational complexity $O(n^{37})$.

In [10], we use quadratic surrogates to make aggregating forecasters more efficient. We show that the resulting algorithm has still high statistical performance for a large class of losses. In particular, we derive an algorithm for multi-class logistic regression with a regret bounded by $O(B\log(n))$ and a computational complexity of only $O(n^4)$.

Online Sign Identification: Minimization of the Number of Errors in Thresholding Bandits

Participants: Réda Ouhamma, Rémy Degenne, Pierre Gaillard, Vianney Perchet.

In the fixed budget thresholding bandit problem, an algorithm sequentially allocates a budgeted number of samples to different distributions. It then predicts whether the mean of each distribution is larger or lower than a given threshold. In [17], we introduce a large family of algorithms (containing most existing relevant ones), inspired by the Frank-Wolfe algorithm, and provide a thorough yet generic analysis of their performance. This allowed us to construct new explicit algorithms, for a broad class

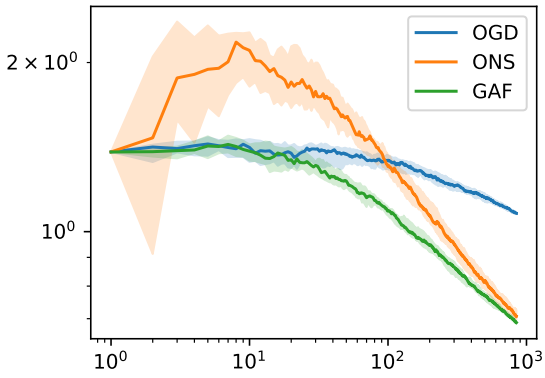

Figure 11: Illustration of the convergence of the proposed algorithm vs. baselines on a multi-class classification dataset.

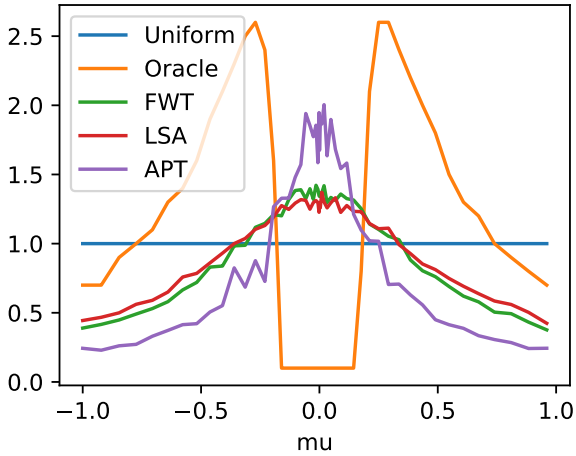

Figure 12: Illustration of the sampling distributions of various strategies and oracles.

of problems, whose losses are within a small constant factor of the non-adaptive oracle ones. Quite interestingly, we observed that adaptive methods empirically greatly out-perform non-adaptive oracles, an uncommon behavior in standard online learning settings, such as regret minimization. We explain this surprising phenomenon on an insightful toy problem.

Dueling Bandits with Adversarial Sleeping

Participants: Aadirupa Saha, Pierre Gaillard.

In [19], we introduce the problem of sleeping dueling bandits with stochastic preferences and adversarial availabilities (DB-SPAA). In almost all dueling bandit applications, the decision space often changes over time; e.g., retail store management, online shopping, restaurant recommendation, search engine optimization, etc. Surprisingly, this 'sleeping aspect' of dueling bandits has never been studied in the literature. Like dueling bandits, the goal is to compete with the best arm by sequentially querying the preference feedback of item pairs. The non-triviality however results due to the non-stationary item spaces that allow any arbitrary subsets items to go unavailable every round. The goal is to find an optimal 'no-regret' policy that can identify the best available item at each round, as opposed to the standard 'fixed best-arm regret objective' of dueling bandits. We first derive an instance-specific lower bound for DB-SPAA $\Omega(\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\log T}{\Delta(i,j)})$, where K is the number of items and $\Delta(i,j)$ is the gap between items i

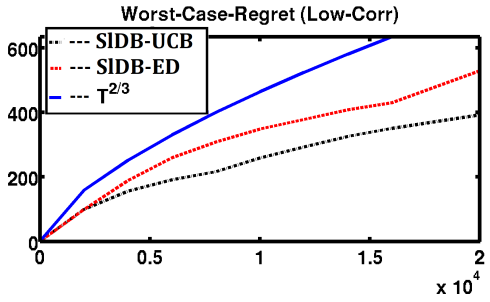

Figure 13: Worst-case regret.

and j. This indicates that the sleeping problem with preference feedback is inherently more difficult than that for classical multi-armed bandits (MAB). We then propose two algorithms, with near optimal regret guarantees. Our results are corroborated empirically.

7.3 Theory and Methods for Deep Neural Networks

A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention

Participants: Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, Julien Mairal.

This work [14] addresses the problem of learning on sets of features, motivated by the need of performing pooling operations in long biological sequences of varying sizes, with long-range dependencies, and possibly few labeled data. To address this challenging task, we introduce a parametrized representation of fixed size, which embeds and then aggregates elements from a given input set according to the optimal transport plan between the set and a trainable reference, see Figure 14. Our approach scales to large datasets and allows end-to-end training of the reference, while also providing a simple unsupervised learning mechanism with small computational cost. Our aggregation technique admits two useful interpretations: it may be seen as a mechanism related to attention layers in neural networks, or it may be seen as a scalable surrogate of a classical optimal transport-based kernel. We experimentally demonstrate the effectiveness of our approach on biological sequences, achieving state-of-the-art results for the protein fold recognition task and detection of chromatin profiles, and, as a proof of concept, we show promising results for processing natural language sequences. We provide an open-source implementation of our embedding that can be used alone or as a module in larger learning models.

GraphiT: Encoding Graph Structure in Transformers.

Participants: Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, Julien Mairal.

In [25], we show that viewing graphs as sets of node features and incorporating structural and positional information into a transformer architecture is able to outperform representations learned with classical graph neural networks (GNNs). Our model, GraphiT, encodes such information by (i) leveraging relative positional encoding strategies in self-attention scores based on positive definite kernels on graphs, and (ii) enumerating and encoding local sub-structures such as paths of short length. We thoroughly evaluate these two ideas on many classification and regression tasks, demonstrating the effectiveness of each of them independently, as well as their combination. In addition to performing well on standard benchmarks, our model also admits natural visualization mechanisms for interpreting graph motifs explaining the predictions, making it a potentially strong candidate for scientific applications where interpretation is important.

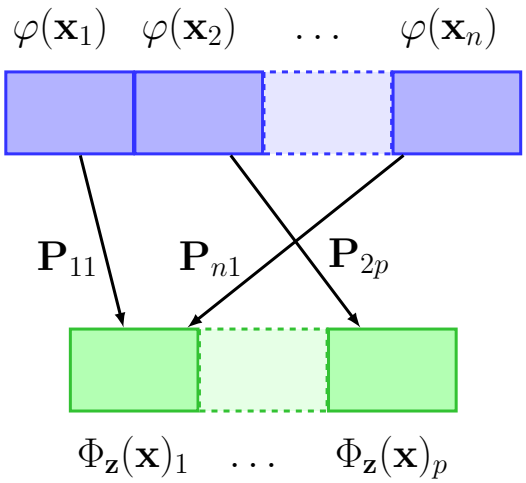

Figure 14: Illustration of our pooling mechanism.

Lucas-Kanade Reloaded: End-to-End Super-Resolution from Raw Image Bursts .

Participants: Bruno Lecouat, Jean Ponce, Julien Mairal.

In [12], we address the problem of reconstructing a high-resolution image from multiple lower-resolution snapshots captured from slightly different viewpoints in space and time. Key challenges for solving this problem include (i) aligning the input pictures with sub-pixel accuracy, (ii) handling raw (noisy) images for maximal faithfulness to native camera data, and (iii) designing/learning an image prior (regularizer) well suited to the task. We address these three challenges with a hybrid algorithm building on the insight from Wronski et al. that aliasing is an ally in this setting, with parameters that can be learned end to end, while retaining the interpretability of classical approaches to inverse problems. The effectiveness of our approach is demonstrated on synthetic and real image bursts, setting a new state of the art on several benchmarks and delivering excellent qualitative results on real raw bursts captured by smartphones and prosumer cameras

A Trainable Spectral-Spatial Sparse Coding Model for Hyperspectral Image Restoration

Participants: Théo Bodrito, Alexandre Zouaoui, Jocelyn Chanussot, Julien Mairal.

Hyperspectral imaging offers new perspectives for diverse applications, ranging from the monitoring of the environment using airborne or satellite remote sensing, precision farming, food safety, planetary exploration, or astrophysics. Unfortunately, the spectral diversity of information comes at the expense of various sources of degradation, and the lack of accurate ground-truth "clean" hyperspectral signals acquired on the spot makes restoration tasks challenging. In particular, training deep neural networks for restoration is difficult, in contrast to traditional RGB imaging problems where deep models tend to shine. In [5], we advocate instead for a hybrid approach based on sparse coding principles that retains the interpretability of classical techniques encoding domain knowledge with handcrafted image priors, while allowing to train model parameters end-to-end without massive amounts of data. We show on various denoising benchmarks that our method is computationally efficient and significantly outperforms the state of the art.

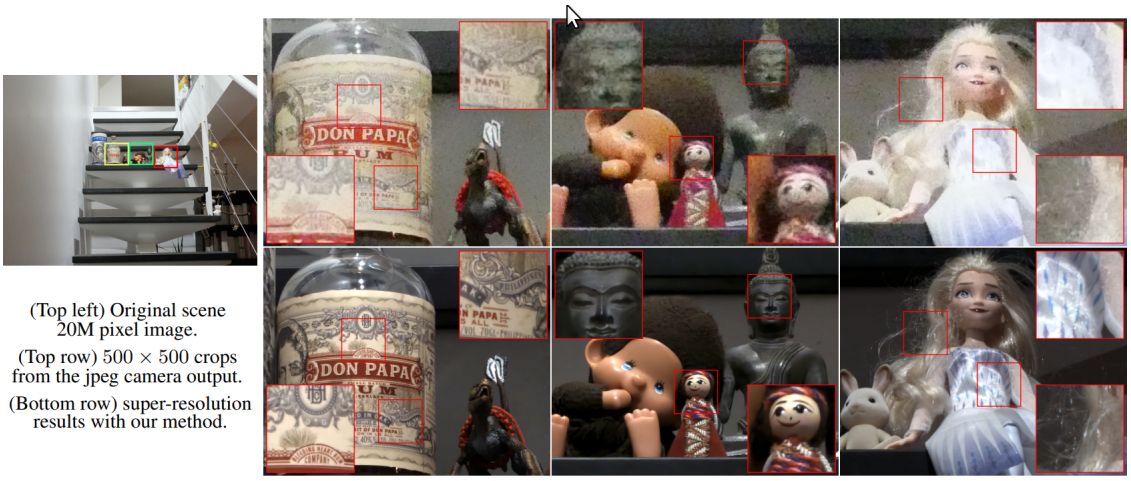

Figure 15: ×4 super-resolution results obtained by processing a burst of 30 raw images acquired with a handheld Panasonic Lumix GX9 camera. Top row: jpeg output of the camera for one frame (high-quality setting); Bottom row: our results with zoomed-in regions. Several bursts of the scene were taken at different ISO settings (12800 for the left image and 25600 for the middle and right images). These are high-noise regimes where small details are lost in the jpeg output.

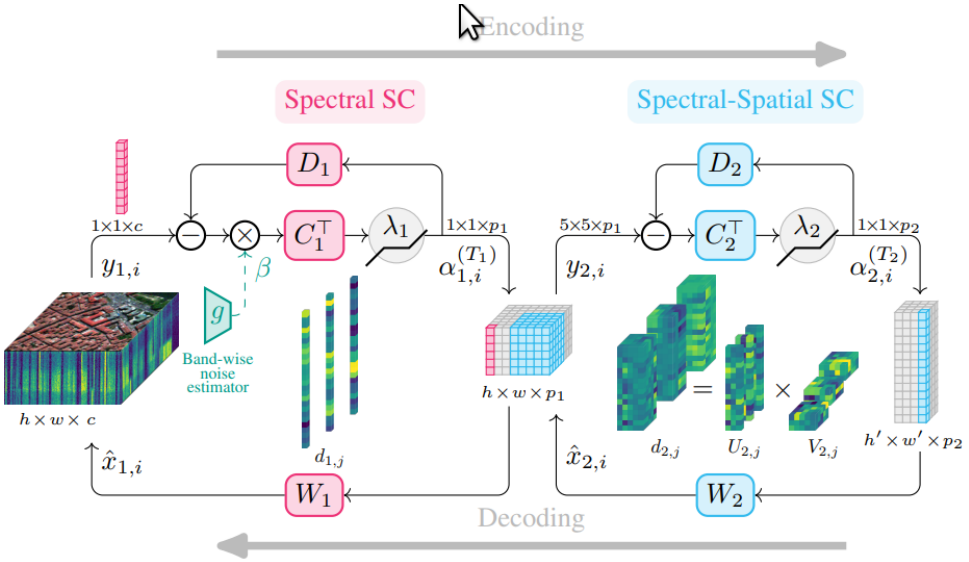

Figure 16: Architecture of T3SC : we propose a two-layer sparse coding model which is end-to-end trainable. The first layer performs a sensor-specific spectral decomposition, while the second layer encodes both spectral and spatial information.

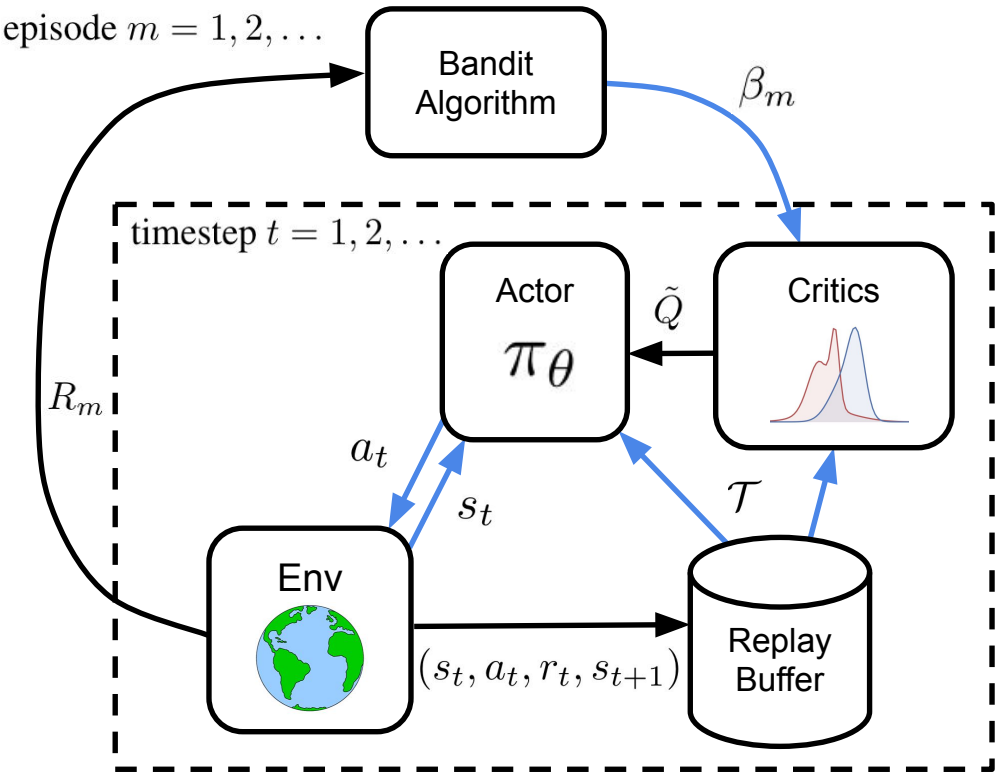

Figure 17: Vizualization of the TOP framework. Blue arrows denote stochastic variables.

Tactical Optimism and Pessimism for Deep Reinforcement Learning

Participants: Ted Moskovitz, Jack Parker-Holder, Aldo Pacchiano, Michael Arbel, Michael I. Jordan.

In recent years, deep off-policy actor-critic algorithms have become a dominant approach to reinforcement learning for continuous control. One of the primary drivers of this improved performance is the use of pessimistic value updates to address function approximation errors, which previously led to disappointing performance. However, a direct consequence of pessimism is reduced exploration, running counter to theoretical support for the efficacy of optimism in the face of uncertainty. So which approach is best? In this work, we show that the most effective degree of optimism can vary both across tasks and over the course of learning. Inspired by this insight, we introduce in [16] a novel deep actor-critic framework, Tactical Optimistic and Pessimistic (TOP) estimation, which switches between optimistic and pessimistic value learning online, see Figure 17. This is achieved by formulating the selection as a multi-arm bandit problem. We show in a series of continuous control tasks that TOP outperforms existing methods which rely on a fixed degree of optimism, setting a new state of the art in challenging pixel-based environments. Since our changes are simple to implement, we believe these insights can easily be incorporated into a multitude of off-policy algorithms.

Towards an Understanding of Default Policies in Multitask Policy Optimization

Participants: Ted Moskovitz, Michael Arbel, Jack Parker-Holder.

Much of the recent success of deep reinforcement learning has been driven by regularized policy optimization (RPO) algorithms, with strong performance across multiple domains. In this family of methods,

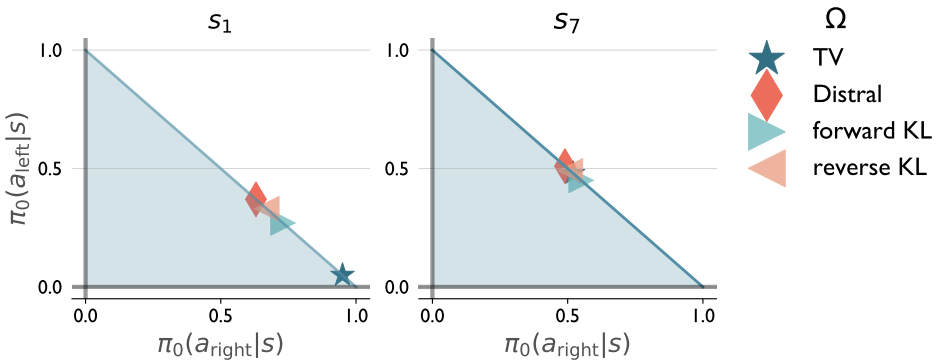

Figure 18: Default policies in two different states at s_1 and s_7 trained on five tasks with a shared structure.

agents are trained to maximize cumulative reward while penalizing deviation in behavior from some reference, or default policy. In addition to empirical success, there is a strong theoretical foundation for understanding RPO methods applied to single tasks, with connections to natural gradient, trust region, and variational approaches. However, there is limited formal understanding of desirable properties for default policies in the multitask setting, an increasingly important domain as the field shifts towards training more generally capable agents. In [15], we take a first step towards filling this gap by formally linking the quality of the default policy to its effect on optimization. Using these results, we then derive a principled RPO algorithm for multitask learning with strong performance guarantees, see Figure 18.

7.4 Pluri-disciplinary Research and Robotics Applications

Memory-Augmented Reinforcement Learning for Image-Goal Navigation

Participants: Lina Mezghani, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, Karteek Alahari.

In this work [24], we address the problem of image-goal navigation in the context of visually-realistic 3D environments. This task involves navigating to a location indicated by a target image in a previously unseen environment. Earlier attempts, including RL-based and SLAM-based approaches, have either shown poor generalization performance, or are heavily reliant on pose/depth sensors. We present a novel method, shown in Figure 19, that leverages a cross-episode memory to learn to navigate. We first train a state-embedding network in a self-supervised fashion, and then use it to embed previously-visited states into an agent's memory. In order to avoid overfitting, we propose to use data augmentation on the RGB input during training. We validate our approach through extensive evaluations, showing that our data-augmented memory-based model establishes a new state of the art on the image-goal navigation task in the challenging Gibson dataset. We obtain this competitive performance from RGB input only, without access to additional sensors such as position or depth.

Episodic Transformer for Vision-and-Language Navigation

Participants: Alexander Pashevich, Cordelia Schmid, Chen Sun.

Interaction and navigation defined by natural language instructions in dynamic environments pose significant challenges for neural agents. This paper [18] focuses on addressing two challenges: handling long sequence of subtasks, and understanding complex human instructions. We propose Episodic Transformer (E.T.), a multimodal transformer that encodes language inputs and the full episode history of

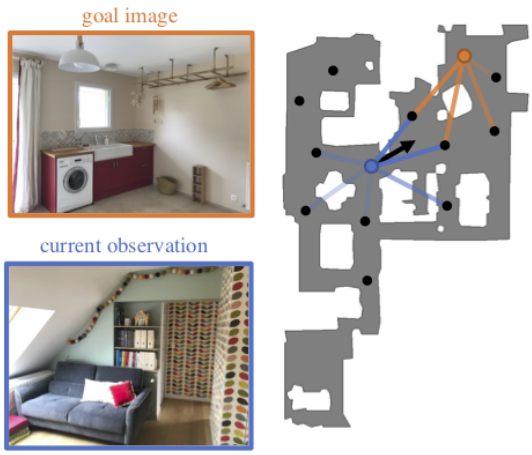

Figure 19: We tackle the problem of image-goal navigation. The agent (shown as the blue dot) is given an image from a goal location (orange dot) which it must navigate to. To address this task, our agent stores a cross-episode memory of previously visited states (black dots), and uses a navigation policy that puts attention (lines) on this memory.

visual observations and actions. To improve training, we leverage synthetic instructions as an intermediate representation that decouples understanding the visual appearance of an environment from the variations of natural language instructions. We demonstrate that encoding the history with a transformer is critical to solve compositional tasks, and that pretraining and joint training with synthetic instructions further improve the performance. Our approach sets a new state of the art on the challenging ALFRED benchmark, achieving 38.4% and 8.5% task success rates on seen and unseen test splits.

Residual Reinforcement Learning from Demonstrations

Participants: Minttu Alakuijala, Gabriel Dulac-Arnold, Julien Mairal, Jean Ponce, Cordelia Schmid.

Residual reinforcement learning (RL) has been proposed as a way to solve challenging robotic tasks by adapting control actions from a conventional feedback controller to maximize a reward signal. In [22], we extend the residual formulation to learn from visual inputs and sparse rewards using demonstrations. Learning from images, proprioceptive inputs and a sparse task-completion reward relaxes the requirement of accessing full state features, such as object and target positions. In addition, replacing the base controller with a policy learned from demonstrations removes the dependency on a hand-engineered controller in favour of a dataset of demonstrations, which can be provided by non-experts. Our experimental evaluation on simulated manipulation tasks on a 6-DoF UR5 arm and a 28-DoF dexterous hand demonstrates that residual RL from demonstrations is able to generalize to unseen environment conditions more flexibly than either behavioral cloning or RL fine-tuning, and is capable of solving high-dimensional, sparse-reward tasks out of reach for RL from scratch.

Extracting Universal Representations of Cognition across Brain-Imaging Studies

Participants: Arthur Mensch, Julien Mairal, Bertrand Thirion, Gael Varoquaux.

We show in [2] how to extract shared brain representations that predict mental processes across many cognitive neuroimaging studies. Focused cognitive-neuroimaging experiments study precise mental processes with carefully-designed cognitive paradigms; however the cost of imaging limits their statistical

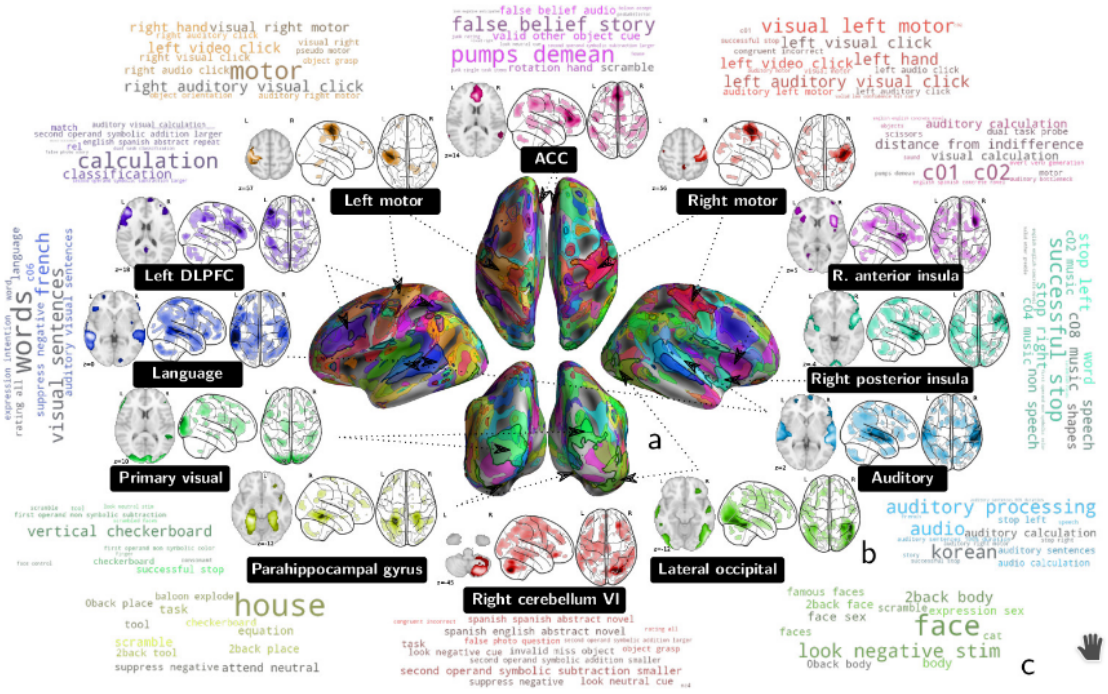

Figure 20: Visualization of some of task-optimized networks. Our approach allows to learn networks that are important for inter-subject decoding across studies. These networks, individually focal and collectively well spread across the cortex, are readily associated with the cognitive tasks that they contribute to predict. We display a selection of these networks, named with the salient anatomical brain region they recruit, along with a word-cloud representation of the stimuli whose likelihood increases with the network activation.

power. On the other hand, large-scale databasing efforts increase considerably the sample sizes, but cannot ask precise cognitive questions. To address this tension, we develop new methods that turn the heterogeneous cognitive information held in different task-fMRI studies into common-universal-cognitive models. Our approach does not assume any prior knowledge of the commonalities shared by the studies in the corpus; those are inferred during model training. The method uses deep-learning techniques to extract representations - task-optimized networks - that form a set of basis cognitive dimensions relevant to the psychological manipulations, as illustrated in Figure 25. In this sense, it forms a novel kind of functional atlas, optimized to capture mental state across many functional-imaging experiments. As it bridges information on the neural support of mental processes, this representation improves decoding performance for 80% of the 35 widely-different functional imaging studies that we consider. Our approach opens new ways of extracting information from brain maps, increasing statistical power even for focused cognitive neuroimaging studies, in particular for those with few subjects.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

Participants: Julien Mairal, Karteek Alahari, Jocelyn Chanussot.

We currently have

- one CIFRE PhD student with Criteo (co-advised by J. Mairal)

- two CIFRE PhD students with Facebook: Mathilde Caron (co-advised by J. Mairal) and Lina Mezghani (co-advised by K. Alahari)
- two CIFRE PhD students with Google: Minttu Alakuijala (co-advised by J. Mairal) and Valentin Gabeur (co-advised by K. Alahari)
- one CIFRE PhD student with Valeo AI: Florent Bartoccioni (co-advised by K. Alahari)
- one CIFRE PhD student with Naver Labs Europe: Bulent Sariyildiz (co-advised by K. Alahari)
- one CIFRE PhD student with Preligens: Jules Bourcier (co-advised by K. Alahari and J. Chanussot)
- one CIFRE PhD student with Nokia Bell Labs: Camila Fernández (co-advised by P. Gaillard)

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

GAYA

Title: Semantic and Geometric Models for Video Interpretation

Duration: 2019 ->

Coordinator: Katerina Fragkiadaki (katef@cs.cmu.edu)

Partners:

- Carnegie Mellon University

Inria contact: Karteek Alahari

Summary: GAYA is a joint research team bringing together two Inria project-teams (Thoth, Grenoble and WILLOW, Paris) and Carnegie Mellon University, USA. It focuses on two research themes: (i) semantic structured interpretation of videos, and (ii) studying the geometric properties of object shapes to enhance state-of-the-art object recognition approaches. More details are available at: Gaya website.

4TUNE

Title: Adaptive, Efficient, Provable and Flexible Tuning for Machine Learning

Duration: 2019 ->

Coordinator: Peter Grünwald (pdg@cwi.nl)

Partners:

- Centrum Wiskunde & Informatica (CWI), Amsterdam

Inria contact: Pierre Gaillard

Summary: The long-term goal of 4TUNE is to push adaptive machine learning to the next level. We aim to develop refined methods, going beyond traditional worst-case analysis, for exploiting structure in the learning problem at hand. We will develop new theory and design sophisticated algorithms for the core tasks of statistical learning and individual sequence prediction. We are especially interested in understanding the connections between these tasks and developing unified methods for both. We will also investigate adaptivity to non-standard patterns encountered in embedded learning tasks, in particular in iterative equilibrium computations. More details are available at: 4TUNE website.

9.2 International research visitors

9.2.1 Visits of international scientists

Other international visits to the team

- Enrico Fini, PhD student from Trento University is visiting us from September 2021 to the end of February 2022.
- Pia Bideau, Postdoctoral researcher from TU Berlin visited us from September to December 2021.

9.3 European initiatives

9.3.1 ERC Starting Grant SOLARIS

Participants: Julien Mairal.

The project SOLARIS started in March 2017 for a duration of five years. The goal of the project is to set up methodological and theoretical foundations of deep learning models, in the context of large-scale data processing. The main applications of the tools developed in this project are for processing visual data, such as videos, but also structured data produced in experimental sciences, such as biological sequences.

The main paradigm used in the project is that of kernel methods and consist of building functional spaces where deep learning models live. By doing so, we want to derive theoretical properties of deep learning models that may explain their success, and also obtain new tools with better stability properties. Another work package of the project is focused on large-scale optimization, which is a key to obtain fast learning algorithms.

9.4 National initiatives

9.4.1 ANR Project AVENUE

Participants: Karteek Alahari.

This ANR project (started in October 2018) aims to address the perception gap between human and artificial visual systems through a visual memory network for human-like interpretation of scenes. To this end, we address three scientific challenges. The first is to learn a network representation of image, video and text data collections, to leverage their inherent diverse cues. The second is to depart from supervised learning paradigms, without compromising on the performance. The third one is to perform inference with the learnt network, e.g., to estimate physical and functional properties of objects, or give cautionary advice for navigating a scene. The principal investigator is Karteek Alahari, and the project involves participants from CentraleSupelec and Ecole des Ponts in Paris.

9.5 Regional initiatives

9.5.1 3IA MIAI chair: Towards More Data Efficiency in Machine Learning

Participants: Julien Mairal, Karteek Alahari, Massih-Reza Amini, Margot Selosse, Juliette Marrie, Romain Ménégaux.

Training deep neural networks when the amount of annotated data is small or in the presence of adversarial perturbations is challenging. More precisely, for convolutional neural networks, it is possible to engineer visually imperceptible perturbations that can lead to arbitrarily different model predictions. Such a robustness issue is related to the problem of regularization and to the ability to generalizing with

few training examples. Our objective is to develop theoretically-grounded approaches that will solve the data efficiency issues of such huge-dimensional models. The principal investigator is Julien Mairal.

10 Dissemination

Participants: Karteek Alahari, Julien Mairal, Pierre Gaillard, Jocelyn Chanussot, Michael Arbel.

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

Member of the organizing committees

- Thoth co-organized the PAISS summer school, an online summer school on AI that has attracted about 300 participants.
- K. Alahari: Program co-chair, BMVC 2021, the premier national conference in computer vision in the UK.
- J. Mairal: co-organizer of the FocusIA workshop at Institut Henri Poincaré, both online and on site. Half a day was targeted to a general audience, in particular high school students.
- J. Mairal: tutorial chair for CVPR 2022. (+10K participants expected).
- J. Mairal: session chair at NeurIPS 2021.
- J. Chanussot: organization of the IEEE Workshop on Hyperspectral Image and Signal Processing:Evolution in Remote Sensing (WHISPERS).
- M. Caron: organization of a tutorial at CVPR 2021.

10.1.2 Scientific events: selection

Member of the conference program committees

- K. Alahari: area chair for CVPR 2021, ICCV 2021, WACV 2022.
- D. Khue Le-Huu: area chair for BMVC 2021.
- J. Mairal: area chair for ICLR 2021 and 2022, AISTATS 2021, ICML 2021 and 2022, and NeurIPS 2021.
- P. Gaillard: area chair for ALT 2021

Reviewer The permanent members, postdocs and senior PhD students of the team reviewed numerous papers for international conferences in artificial intelligence, computer vision and machine learning, including ACM Multimedia, AISTATS, CVPR, ICCV, ICML, ICLR, COLT, ALT, NeurIPS in 2021.

10.1.3 Journal

Member of the editorial boards

- K. Alahari: Associate editor of the International Journal of Computer Vision, since 2019.
- K. Alahari: Associate editor of the Computer Vision and Image Understanding journal, since 2018.
- J. Mairal: Associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), since 2021.

- J. Mairal: Associate editor of the Journal of Machine Learning Research (JMLR), since 2019.
- J. Mairal: Associate editor of the International Journal of Computer Vision, since 2015.
- J. Mairal: Associate editor of Journal of Mathematical Imaging and Vision, since 2015.

Reviewer - reviewing activities The permanent members, postdocs and senior PhD students of the team reviewed numerous papers for international journals in computer vision (IJCV, PAMI, CVIU), machine learning (JMLR).

10.1.4 Invited talks

- M. Arbel: invited seminar at Instituto Superior Técnico (online).
- M. Arbel: invited talk at "Les journées MAS, Institut Denis Poisson"
- M. Arbel: invited seminar at Flatiron Institute/NYU (online).
- M. Arbel: invited seminar at Ecole Normale Supérieure, Lyon.
- M. Arbel: invited seminar at Laboratoire Jean Kuntzmann, Grenoble.
- H. Leterme: CollaboTICS - International Collaborative Workshop of RUB-UGA-UT, 3rd Edition (online).
- J. Mairal: invited talk at the French-German Machine Learning Symposium, Munich (online).
- J. Mairal: invited seminar at Flatiron Institute/NYU (online).
- J. Mairal: invited seminar at KU Leuven (online).
- J. Mairal: invited talk at the Prairie workshop. Paris.
- J. Mairal: keynote speaker at the Mascot NUM conference. (online).
- P. Gaillard: invited seminar at LJK. Grenoble.
- P. Gaillard: invited seminar at the Aptikal research seminar. Grenoble.

10.1.5 Scientific expertise

- K. Alahari: reviewer for ANR.
- J. Mairal: reviewer for the Swiss Data Science Center.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Doctorat: K. Alahari, Lecturer at the Machine Learning Summer School (MLSS), 3h eqTD, Taipei, Taiwan (online).
- Doctorat: K. Alahari, Lecturer at the CVIT summer school on machine learning, 3h eqTD, IIIT Hyderabad, India (online).
- Master: K. Alahari, Machine Learning for Multimodal Data, 11.25h eqTD, M2, UGA, Grenoble.
- Master: K. Alahari, Understanding Big Visual Data, 13.5h eqTD, M2, Grenoble INP.
- Master: K. Alahari, Graphical Models Inference and Learning, 18h eqTD, M2, CentraleSupelec, Paris.
- Master: K. Alahari, Introduction to computer vision, 9h eqTD, M1, ENS Paris.

- Master: H. Leterme, Analyse pour l'ingénieur, 31.5h eqTD, Grenoble INP Ensimag.
- Master: H. Leterme, Traitement d'images, 12h eqTD, Grenoble INP Ensimag.
- Master: J. Mairal, Kernel methods for statistical learning, 17h eqTD, M2, Ecole Normale Supérieure, Cachan, France.
- Master: J. Mairal, Kernel methods for statistical learning, 27h eqTD, M2, UGA, Grenoble.
- Master: P. Gaillard, Sequential Learning, 27h eqTD, M2, Ecole Normale Supérieure, Cachan, France.

10.2.2 Supervision (PhD defenses)

- PhD: Vladyslav Sydorov: Techniques spatiales pour la compréhension vidéo. 10/05/2021. Dir: Karteek Alahari, Cordelia Schmid.
- PhD: Mathilde Caron: Apprentissage auto-supervisé de représentations visuelles avec des réseaux de neurones profonds. 09/12/2021. Dir: Julien Mairal.
- PhD: Alexander Pashevich: Des robots qui voient : Apprentissage de comportements guidés par la vision. Univ. Grenoble Alpes. 29/09/2021. Dir: Cordelia Schmid.
- PhD: Raphaël Berthier: Analysis and Acceleration of Gradient Descents and Gossip Algorithms. 25/09/2021. Dir: Pierre Gaillard.

10.2.3 Juries

- K. Alahari: Reviewer for the PhD thesis of Enzo Battistella, Université Paris-Saclay.
- K. Alahari: Jury member of the thesis defense committee of Allison Del Giorno, Carnegie Mellon University.
- J. Mairal: Reviewer for the PhD thesis of Tran Khanh Hung, Université Paris-Saclay.
- J. Mairal: Reviewer for the PhD thesis of Hamza Cherkaoui, Université Paris-Saclay.
- J. Mairal: Reviewer for the PhD thesis of Thomas Eboli, PSL-Sorbonne Université.
- J. Mairal: Jury member for the HdR of Chaohui Wang, Université Paris-Est.
- J. Mairal: Jury member for the HdR of Franck Iutzeler, Université Grenoble-Alpes.
- J. Mairal: CSI member for the PhD of Tayeb Zarrouk. Univ. Grenoble Alpes.

10.3 Popularization

10.3.1 Interventions

- J. Mairal was interviewed for a podcast from IHP called "l'Oreille Mathématique".

11 Scientific production

11.1 Publications of the year

International journals

[1] A. Bietti, A. Agarwal and J. Langford. 'A Contextual Bandit Bake-off'. In: *Journal of Machine Learning Research* 22.133 (2021), pp. 1–49. URL: https://hal.inria.fr/hal-01708310.

[2] A. Mensch, J. Mairal, B. Thirion and G. Varoquaux. 'Extracting representations of cognition across neuroimaging studies improves brain decoding'. In: *PLoS Computational Biology* 17.5 (3rd May 2021), e1008795:1–20. DOI: 10.1371/journal.pcbi.1008795. URL: https://hal.archives-ouvertes.fr/hal-01874713.

International peer-reviewed conferences

[3] M. Arbel and J. Mairal. 'Amortized implicit differentiation for stochastic bilevel optimization'. In: The Tenth International Conference on Learning Representations. Online, France, 2022. URL: https://hal.archives-ouvertes.fr/hal-03455458.

[4] G. Beugnot, J. Mairal and A. Rudi. 'Beyond Tikhonov: Faster Learning with Self-Concordant Losses via Iterative Regularization'. In: NeurIPS 2021 – 35th Annual Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems 34. Virtual, France, 6th Dec. 2021, pp. 1–37. URL: https://hal.inria.fr/hal-03406072.

[5] T. Bodrito, A. Zouaoui, J. Chanussot and J. Mairal. 'A Trainable Spectral-Spatial Sparse Coding Model for Hyperspectral Image Restoration'. In: NeurIPS 2021 – 35th Annual Conference on Neural Information Processing Systems. Sydney, Australia, 6th Dec. 2021, pp. 1–19. URL: https://hal.archives-ouvertes.fr/hal-03423559.

[6] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski and A. Joulin. 'Emerging Properties in Self-Supervised Vision Transformers'. In: ICCV 2021 - International Conference on Computer Vision. Virtual, France, 11th Oct. 2021, pp. 1–21. URL: https://hal.archives-ouvertes.fr/hal-03323359.

[7] M. Even, R. Berthier, F. Bach, N. Flammarion, P. Gaillard, H. Hendrikx, L. Massoulié and A. Taylor. 'A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip'. In: *Advances in Neural Information Processing Systems 34*. NeurIPS 2021 - 35th Conference on Neural Information Processing Systems. Sydney (virtual), Australia: Morgan Kaufmann Publishers, 1st Dec. 2021, pp. 1–32. URL: https://hal.archives-ouvertes.fr/hal-03405165.

[8] C. Fernández, C. Shue Chen, P. Gaillard and A. Silva. 'Experimental Comparison of Semi-parametric, Parametric, and Machine Learning Methods for Time-to-Event Analysis Through the IPEC Score'. In: SFdS 2020 - 52èmes Journées de Statistiques de la Société Française de Statistique. Nice, France, 7th June 2021, pp. 1–6. URL: https://hal.inria.fr/hal-03221512.

[9] P. Glaser, M. Arbel and A. Gretton. 'KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support'. In: NeurIPS 2021 -Thirty-Fifth Annual Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems. Online, France, 2021, pp. 1–29. URL: https://hal.archives-ouvertes.fr/hal-03455473.

[10] R. Jézéquel, P. Gaillard and A. Rudi. 'Mixability made efficient: Fast online multiclass logistic regression'. In: NeurIPS 2021. Thirty-fifth Conference on Neural Information Processing Systems. Online, France, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03370530.

[11] Đ. Lê-Huu and K. Alahari. 'Regularized Frank-Wolfe for Dense CRFs: Generalizing Mean Field and Beyond'. In: NeurIPS 2021 - 35th Annual Conference on Neural Information Processing Systems. Virtual-only Conference, Australia, 6th Dec. 2021, pp. 1–35. URL: https://hal.inria.fr/hal-03406107.

[12] B. Lecouat, J. Ponce and J. Mairal. 'Lucas-Kanade Reloaded: End-to-End Super-Resolution from Raw Image Bursts'. In: ICCV 2021 - International Conference on Computer Vision. Virtual, France, 2021, pp. 1–16. URL: https://hal.inria.fr/hal-03323885.

[13] H. Leterme, K. Polisano, V. Perrier and K. Alahari. 'Modélisation Parcimonieuse de CNNs avec des Paquets d'Ondelettes Dual-Tree'. In: ORASIS 2021 - Journées francophones des jeunes chercheurs en vision par ordinateur. Saint Ferréol, France, 13th Sept. 2021, pp. 1–9. URL: https://hal.archives-ouvertes.fr/hal-03339792.

[14] G. Mialon, D. Chen, A. D'Aspremont and J. Mairal. 'A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention'. In: ICLR 2021 - The Ninth International Conference on Learning Representations. Virtual, France, 4th May 2021. URL: https://hal.archives-ouvertes.fr/hal-02883436.

[15] T. Moskovitz, M. Arbel, J. Parker-Holder and A. Pacchiano. 'Towards an Understanding of Default Policies in Multitask Policy Optimization'. In: 25th International Conference on Artificial Intelligence and Statistics. Volume 130: International Conference on Artificial Intelligence and Statistics. Online, France, 2022. URL: https://hal.archives-ouvertes.fr/hal-03455465.

[16] T. Moskovitz, J. Parker-Holder, A. Pacchiano, M. Arbel and M. I. Jordan. 'Tactical Optimism and Pessimism for Deep Reinforcement Learning'. In: NeurIPS 2021 - Thirty-fifth Annual Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems. Online, France, 2021, pp. 1–15. URL: https://hal.archives-ouvertes.fr/hal-03455481.

[17] R. Ouhamma, R. Degenne, P. Gaillard and V. Perchet. 'Online Sign Identification: Minimization of the Number of Errors in Thresholding Bandits'. In: NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. Virtual, Canada, 2021, pp. 1–25. URL: https://hal.inria.fr/hal-03363014.

[18] A. Pashevich, C. Schmid and C. Sun. 'Episodic Transformer for Vision-and-Language Navigation'. In: ICCV 2021 - International Conference on Computer Vision. Virtual, United States, 11th Oct. 2021, pp. 1–18. URL: https://hal.inria.fr/hal-03371803.

[19] A. Saha and P. Gaillard. 'Dueling Bandits with Adversarial Sleeping'. In: NeurIPS 2021 - 35th International Conference on Neural Information Processing Systems. Virtual, Canada, 6th Dec. 2021, pp. 1–25. URL: https://hal.inria.fr/hal-03451845.

[20] M. B. Sariyildiz, Y. Kalantidis, D. Larlus and K. Alahari. 'Concept Generalization in Visual Representation Learning'. In: ICCV 2021 - International Conference on Computer Vision. Virtual, Canada, 11th Oct. 2021. URL: https://hal.inria.fr/hal-03110632.

Conferences without proceedings

[21] V. Gabeur, A. Nagrani, C. Sun, K. Alahari and C. Schmid. 'Masking Modalities for Cross-modal Video Retrieval'. In: WACV 2022 - Winter Conference on Applications of Computer Vision. Waikoloa, United States, 4th Jan. 2022, pp. 1–10. URL: https://hal.archives-ouvertes.fr/hal-03420133.

Reports & preprints

[22] M. Alakuijala, G. Dulac-Arnold, J. Mairal, J. Ponce and C. Schmid. *Residual Reinforcement Learning from Demonstrations*. 15th June 2021. URL: https://hal.inria.fr/hal-03260683.

[23] F. Bartoccioni, É. Zablocki, P. Pérez, M. Cord and K. Alahari. *LiDARTouch: Monocular metric depth estimation with a few-beam LiDAR*. Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03508099.

[24] L. Mezghani, S. Sukhbaatar, T. Lavril, O. Maksymets, D. Batra, P. Bojanowski and K. Alahari. *Memory-Augmented Reinforcement Learning for Image-Goal Navigation*. 14th Jan. 2021. URL: https://hal.inria.fr/hal-03110875.

[25] G. Mialon, D. Chen, M. Selosse and J. Mairal. *GraphiT: Encoding Graph Structure in Transformers*. 10th June 2021. URL: https://hal.archives-ouvertes.fr/hal-03256708.

[26] H. Zenati, A. Bietti, M. Martin, E. Diemert and J. Mairal. *Counterfactual Learning of Stochastic Policies with Continuous Actions: from Models to Offline Evaluation*. 19th Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-02883423.