Activity Report 2019

Project-Team THOTH

Learning visual models from large-scale data

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Vision, perception and multimedia interpretation

Table of contents

Project-Team THOTH

Creation of the Team: 2016 January 01, updated into Project-Team: 2016 March 01

Keywords:

Computer Science and Digital Science:

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:

B5.6. - Robotic systems
B8.4. - Security and personal assistance
B8.5. - Smart society
B9.5.1. - Computer science
B9.5.6. - Data science

1. Team, Visitors, External Collaborators

Research Scientists

Julien Mairal [Team leader, Inria, Researcher, HDR]
Cordelia Schmid [Inria, Senior Researcher, HDR]
Karteek Alahari [Inria, Researcher, HDR]
Grégory Rogez [Inria, Starting Research Position, until Jan 2019]
Jakob Verbeek [Inria, Senior Researcher, HDR]

Faculty Member

Jocelyn Chanussot [Institut polytechnique de Grenoble, Professor, from Sep 2019, HDR]

Post-Doctoral Fellow

Adria Ruiz Ovejero [Inria, Post-Doctoral Fellow]

PhD Students

Minttu Alakuijala [Inria, PhD Student]
Florent Bartoccioni [Inria, PhD Student, from Nov 2019]
Alberto Bietti [Inria, PhD Student]
Mathilde Caron [Facebook, PhD Student, granted by CIFRE]
Dexiong Chen [Univ. Grenoble Alpes, PhD Student]
Mikita Dvornik [Inria, PhD Student]
Maha Elbayad [Univ Grenoble Alpes, PhD Student]
Valentin Gabeur [Inria, PhD Student]
Pierre Louis Guhur [Univ Paris-Saclay, PhD Student, from Sep 2019]
Yana Hasson [Inria, PhD Student]
Ekaterina Iakovleva [Univ Grenoble Alpes, PhD Student]

Roman Klokov [Inria, PhD Student]
Andrei Kulunchakov [Inria, PhD Student]
Bruno Lecouat [Inria, PhD Student, from Sep 2019]
Hubert Leterme [Univ Grenoble Alpes, PhD Student, from Sep 2019]
Pauline Luc [Facebook CIFRE, PhD Student, until May 2019]
Thomas Lucas [Univ Grenoble Alpes, PhD Student]
Lina Mezghani [Inria, PhD Student, from Nov 2019]
Gregoire Mialon [Inria, PhD Student]
Alexander Pashevich [Inria, PhD Student]
Alexandre Sablayrolles [Facebook CIFRE, PhD Student]
Konstantin Shmelkov [Inria, PhD Student, until Mar 2019]
Robin Strudel [École Normale Supérieure de Paris, PhD Student]
Vladyslav Sydorov [Inria, PhD Student]
Gul Varol Simsekli [Inria, PhD Student, until Feb 2019]
Nitika Verma [Univ Grenoble Alpes, PhD Student, until May 2019]
Daan Wynen [Inria, PhD Student, until Sep 2019]
Houssam Zenati [Criteo, PhD Student, from Oct 2019]

Technical staff

Ghislain Durif [Inria, Engineer, until Jan 2019]
Ricardo Jose Garcia Pinel [Inria, Engineer, from Jul 2019]
François Gindraud [Inria, Engineer, until Mar 2019]
Igor Kalevatykh [Inria, Engineer, until Nov 2019]
Bruno Lecouat [Inria, Engineer, until Aug 2019]
Xavier Martin [Inria, Engineer]
Alexandre Zouaoui [Inria, Engineer, from Dec 2019]

Interns and Apprentices

Pierre Louis Guhur [Univ Paris-Saclay, from Apr 2019 until Aug 2019]
Hubert Leterme [Univ Grenoble Alpes, from Feb 2019 until Jul 2019]
Matthieu Toulemont [ENPC Paris, from Apr 2019 until Sep 2019]
Houssam Zenati [Criteo, from Apr 2019 until Oct 2019]

Administrative Assistant

Nathalie Gillot [Inria, Administrative Assistant]

Visiting Scientists

Hernan Dario Benitez Restrepo [Pontificia Universidad Javeriana Sede Cali, from Nov 2019]
Pia Bideau [Univ. Massachusetts Amherst, until Jan 2019]
Avijit Dasgupta [IIIT Hyderabad, from Feb 2019 until May 2019]
Ning Huyan [from Dec 2019]
Dou Quan [Inria, from Dec 2019]
Gunnar Atli Sigurdsson [CMU, until Mar 2019]

External Collaborator

Ghislain Durif [CNRS, from Feb 2019]

2. Overall Objectives

2.1. Overall Objectives

In 2021, it is expected that nearly 82% of the Internet traffic will be due to videos, and that it would take an individual over 5 million years to watch the amount of video that will cross global IP networks each month by then. Thus, there is a pressing and in fact increasing demand to annotate and index this visual content for home and professional users alike. The available text and speech-transcript metadata is typically not sufficient by itself for answering most queries, and visual data must come into play. On the other hand, it is not imaginable to learn the models of visual content required to answer these queries by manually and precisely annotating every relevant concept, object, scene, or action category in a representative sample of everyday conditions—if only because it may be difficult, or even impossible to decide a priori what are the relevant categories and the proper granularity level. This suggests reverting back to the original metadata as source of annotation, despite the fact that the information it provides is typically sparse (e.g., the location and overall topic of newscasts in a video archive) 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 screen or not have survived the final cut). On the other hand, this weak form of "embedded annotation" is rich and diverse, and mining the corresponding visual data from the web, TV or film archives guarantees that it is representative of the many different scene settings depicted in situations typical of on-line content. Thus, leveraging this largely untapped source of information, rather than attempting to hand label all possibly relevant visual data, is a key to the future use of on-line imagery.

Today's object recognition and scene understanding technology operates in a very different setting; it mostly relies on fully supervised classification engines, and visual models are essentially (piecewise) rigid templates learned from hand labeled images. The sheer scale of on-line data and the nature of the embedded annotation call for a departure from this fully supervised scenario. The main idea of the Thoth project-team is to develop a new framework for learning the structure and parameters of visual models by actively exploring large digital image and video sources (off-line archives as well as growing on-line content, with millions of images and thousands of hours of video), and exploiting the weak supervisory signal provided by the accompanying metadata. This huge volume of visual training data will allow us to learn complex non-linear models with a large number of parameters, such as deep convolutional networks and higher-order graphical models. This is an ambitious goal, given the sheer volume and intrinsic variability of the visual data 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. Further, recent advances at a smaller scale suggest that this is realistic. For example, it is already possible to determine the identity of multiple people from news images and their captions, or to learn human action models from video scripts. There has also been recent progress in adapting supervised machine learning technology to large-scale settings, where the training data is very large and potentially infinite, and some of it may not be labeled. Methods that adapt the structure of visual models to the data are also emerging, and the growing computational power and storage capacity of modern computers are enabling factors that should of course not be neglected.

One of the main objective of Thoth is to transform massive visual data into trustworthy knowledge libraries. For that, it addresses several challenges.

- designing and learning structured models capable of representing complex visual information.
- learning visual models from minimal supervision or unstructured meta-data.
- large-scale learning and optimization.

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, estimating human poses, recovering scene geometry, recognizing activities performed by humans. 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, people on a road are usually walking or standing, (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 three 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 focus of the second topic is the challenging problem of modeling human activities in video, starting from human activity descriptors to building intermediate spatio-temporal representations of videos, and then learning the interactions among humans, objects and scenes temporally. The last 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 the detection of people and their body-joint locations in video, 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. For the application of recognizing human activities, this involves learning deep features for humans and their body-parts with all their spatiotemporal variations, either directly from raw video data or "pre-processed" videos containing human detections. For the application of object tracking, this task amounts to learning object-specific deep representations, further exploiting the limited annotation provided to identify the object.
- **Modeling human activities in videos.** Humans and their activities are not only one of the most frequent and interesting subjects in videos but also one of the hardest to analyze owing to the complexity of the human form, clothing and movements. As part of this task, the Thoth project-team plans to build on state-of-the-art approaches for spatio-temporal representation of videos. This will involve using the dominant motion in the scene as well as the local motion of individual parts undergoing a rigid motion. Such motion information also helps in reasoning occlusion relationships among people and objects, and the state of the object. This novel spatio-temporal representation ultimately provides the equivalent of object proposals for videos, and is an important component for learning algorithms using minimal supervision. To take this representation even further, we aim to integrate the proposals and the occlusion relationships with methods for estimating human pose in videos, thus leveraging the interplay among body-joint locations, objects in the scene, and the activity

being performed. For example, the locations of shoulder, elbow and wrist of a person drinking coffee are constrained to move in a certain way, which is completely different from the movement observed when a person is typing. In essence, this step will model human activities by dynamics in terms of both low-level movements of body-joint locations and global high-level motion in the scene.

- **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 video, e.g., a prior 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 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

[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.

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 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 amount 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 labelled 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.

3.4. Datasets and evaluation

Standard benchmarks with associated evaluation measures are becoming increasingly important in computer vision, as they enable an objective comparison of state-of-the-art approaches. Such datasets need to be relevant for real-world application scenarios; challenging for state-of-the-art algorithms; and large enough to produce statistically significant results.

A decade ago, small datasets were used to evaluate relatively simple tasks, such as for example interest point matching and detection. Since then, the size of the datasets and the complexity of the tasks gradually evolved. An example is the Pascal Visual Object Challenge with 20 classes and approximately 10,000 images, which evaluates object classification and detection. Another example is the ImageNet challenge, including thousands of classes and millions of images. In the context of video classification, the TrecVid Multimedia Event Detection challenges, organized by NIST, evaluate activity classification on a dataset of over 200,000 video clips, representing more than 8,000 hours of video, which amounts to 11 months of continuous video.

Almost all of the existing image and video datasets are annotated by hand; it is the case for all of the above cited examples. In some cases, they present limited and unrealistic viewing conditions. For example, many images of the ImageNet dataset depict upright objects with virtually no background clutter, and they may not capture particularly relevant visual concepts: most people would not know the majority of subcategories of snakes cataloged in ImageNet. This holds true for video datasets as well, where in addition a taxonomy of action and event categories is missing.

Our effort on data collection and evaluation will focus on two directions. First, we will design and assemble video datasets, in particular for action and activity recognition. This includes defining relevant taxonomies of actions and activities. Second, we will provide data and define evaluation protocols for weakly supervised learning methods. This does not mean of course that we will forsake human supervision altogether: some amount of ground-truth labeling is necessary for experimental validation and comparison to the state of the art. Particular attention will be payed to the design of efficient annotation tools.

Not only do we plan to collect datasets, but also to provide them to the community, together with accompanying evaluation protocols and software, to enable a comparison of competing approaches for action recognition and large-scale weakly supervised learning. Furthermore, we plan to set up evaluation servers together with leaderboards, to establish an unbiased state of the art on held out test data for which the ground-truth annotations are not distributed. This is crucial to avoid tuning the parameters for a specific dataset and to guarantee a fair evaluation.

- **Action recognition.** We will develop datasets for recognizing human actions and human-object interactions (including multiple persons) with a significant number of actions. Almost all of today's action recognition datasets evaluate classification of short video clips into a number of predefined categories, in many cases a number of different sports, which are relatively easy to identify by their characteristic motion and context. However, in many real-world applications the goal is to identify and localize actions in entire videos, such as movies or surveillance videos of several hours. The actions targeted here are "real-world" and will be defined by compositions of atomic actions into higher-level activities. One essential component is the definition of relevant taxonomies of actions and activities. We think that such a definition needs to rely on a decomposition of actions into poses, objects and scenes, as determining all possible actions without such a decomposition is not feasible. We plan to provide annotations for spatio-temporal localization of humans as well as relevant objects and scene parts for a large number of actions and videos.
- **Weakly supervised learning.** We will collect weakly labeled images and videos for training. The collection process will be semi-automatic. We will use image or video search engines such as Google Image Search, Flickr or YouTube to find visual data corresponding to the labels. Initial datasets will be obtained by manually correcting whole-image/video labels, i.e., the approach will evaluate how well the object model can be learned if the entire image or video is labeled, but the object model has to be extracted automatically. Subsequent datasets will features noisy and incorrect labels. Testing will be performed on PASCAL VOC'07 and ImageNet, but also on more realistic datasets similar

to those used for training, which we develop and manually annotate for evaluation. Our dataset will include both images and videos, the categories represented will include objects, scenes as well as human activities, and the data will be presented in realistic conditions.

- **Joint learning from visual information and text.** Initially, we will use a selection from the large number of movies and TV series for which scripts are available on-line, see for example http://www.dailyscript.com and http://www.weeklyscript.com. These scripts can easily be aligned with the videos by establishing correspondences between script words and (timestamped) spoken ones obtained from the subtitles or audio track. The goal is to jointly learn from visual content and text. To measure the quality of such a joint learning, we will manually annotate some of the videos. Annotations will include the space-time locations of the actions as well as correct parsing of the sentence. While DVDs will, initially, receive most attention, we will also investigate the use of data obtained from web pages, for example images with captions, or images and videos surrounded by text. This data is by nature more noisy than scripts.

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 surveillance, service robotics for assistance in day-to-day activities as well as the medical domain.
- Action recognition is highly relevant to visual surveillance, assisted driving and video access.
- 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. In particular,

- extensions of unsupervised learning techniques originally developed for modelling the statistics of natural images have been deployed in neuro-imaging for fMRI data with the collaboration of the Parietal team from Inria.
- similarly, deep convolutional data representations, also originally developed for visual data, have been successfully extended to the processing of biological sequences, with collaborators from bio-informatics.
- Thoth also collaborates with experts in natural language and text processing, for applications where visual modalities need to be combined with text data.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

- Cordelia Schmid received the Royal Society Milner Award, 2019.
- Julien Mairal received the test-of-time award at the International Conference on Machine Learning (ICML), 2019.
- The paper [21] authored by Roman Klokov, Jakob Verbeek, Edmond Boyer [Inria Morpheo] won the “Best Science Paper Award Honourable Mention” at BMVC 2019.
- Jakob Verbeek was awarded as an outstanding reviewer at ICLR 2019.
- Adria Ruiz Ovejero was awarded as an outstanding reviewer at ICCV 2019.

5.1.2. Dissemination

- The team co-organized PAISS 2019, an international AI summer school in Paris. This is the second edition of the school that was first organized in Grenoble in 2018. The 2019 edition brought together over 200 participants. We also provided scholarships to 21 students to encourage diversity among the attendees.

6. New Software and Platforms

6.1. LCR-Net

Localization-Classification-Regression Network for Human Pose

KEYWORDS: Object detection - Recognition of human movement

FUNCTIONAL DESCRIPTION: We propose an end-to-end architecture for joint 2D and 3D human pose estimation in natural images. Key to our approach is the generation and scoring of a number of pose proposals per image, which allows us to predict 2D and 3D pose of multiple people simultaneously. Our architecture contains 3 main components: 1) the pose proposal generator that suggests potential poses at different locations in the image, 2) a classifier that scores the different pose proposals , and 3) a regressor that refines pose proposals both in 2D and 3D.

- Participants: Grégory Rogez, Philippe Weinzaepfel and Cordelia Schmid
- Partner: Naver Labs Europe
- Contact: Nicolas Jourdan
- Publication: LCR-Net: Localization-Classification-Regression for Human Pose
- URL: https://thoth.inrialpes.fr/src/LCR-Net/

6.2. CKN-seq

Convolutional Kernel Networks for Biological Sequences

KEYWORD: Bioinformatics

SCIENTIFIC DESCRIPTION: The growing amount of biological sequences available makes it possible to learn genotype-phenotype relationships from data with increasingly high accuracy. By exploiting large sets of sequences with known phenotypes, machine learning methods can be used to build functions that predict the phenotype of new, unannotated sequences. In particular, deep neural networks have recently obtained good performances on such prediction tasks, but are notoriously difficult to analyze or interpret. Here, we introduce a hybrid approach between kernel methods and convolutional neural networks for sequences, which retains the ability of neural networks to learn good representations for a learning problem at hand, while defining a well characterized Hilbert space to describe prediction functions. Our method outperforms state-of-the-art convolutional neural networks on a transcription factor binding prediction task while being much faster to train and yielding more stable and interpretable results.

FUNCTIONAL DESCRIPTION: D. Chen, L. Jacob, and J. Mairal. Biological Sequence Modeling with Convolutional Kernel Networks. Bioinformatics, volume 35, issue 18, pages 3294-3302, 2019.

- Participants: Laurent Jacob, Dexiong Chen and Julien Mairal
- Partners: CNRS - UGA
- Contact: Julien Mairal
- Publication: Biological Sequence Modeling with Convolutional Kernel Networks
- URL: https://gitlab.inria.fr/dchen/CKN-seq

6.3. LVO

Learning Video Object Segmentation with Visual Memory

KEYWORD: Video analysis

FUNCTIONAL DESCRIPTION: This is a public implementation of the method described in the following paper: Learning Video Object Segmentation with Visual Memory [ICCV 2017] (https://hal.archives-ouvertes.fr/hal-01511145v2/document).

This paper addresses the task of segmenting moving objects in unconstrained videos. We introduce a novel two-stream neural network with an explicit memory module to achieve this. The two streams of the network encode spatial and temporal features in a video sequence respectively, while the memory module captures the evolution of objects over time. The module to build a "visual memory" in video, i.e., a joint representation of all the video frames, is realized with a convolutional recurrent unit learned from a small number of training video sequences. Given a video frame as input, our approach assigns each pixel an object or background label based on the learned spatio-temporal features as well as the "visual memory" specific to the video, acquired automatically without any manually-annotated frames. The visual memory is implemented with convolutional gated recurrent units, which allows to propagate spatial information over time. We evaluate our method extensively on two benchmarks, DAVIS and Freiburg-Berkeley motion segmentation datasets, and show state-of-the-art results. For example, our approach outperforms the top method on the DAVIS dataset by nearly 6%. We also provide an extensive ablative analysis to investigate the influence of each component in the proposed framework.

- Participants: Karteek Alahari, Cordelia Schmid and Pavel Tokmakov
- Contact: Pavel Tokmakov
- Publication: hal-01511145v2
- URL: http://lear.inrialpes.fr/research/lvo/

6.4. SURREAL

Learning from Synthetic Humans

KEYWORDS: Synthetic human - Segmentation - Neural networks

FUNCTIONAL DESCRIPTION: The SURREAL dataset consisting of synthetic videos of humans, and models trained on this dataset are released in this package. The code for rendering synthetic images of people and for training models is also included in the release.

- Participants: Gül Varol, Xavier Martin, Ivan Laptev and Cordelia Schmid
- Contact: Gül Varol
- Publication: Learning from Synthetic Humans
- URL: http://www.di.ens.fr/willow/research/surreal/

6.5. attn2d

Pervasive Attention

KEYWORDS: NLP - Deep learning - Machine translation

SCIENTIFIC DESCRIPTION: Pervasive attention : 2D Convolutional Networks for Sequence-to-Sequence Prediction

FUNCTIONAL DESCRIPTION: An open source PyTorch implementation of the pervasive attention model described in: Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2018. Pervasive Attention: 2D Convolutional Networks for Sequence-to-Sequence Prediction. In Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018)

- Participants: Maha Elbayad and Jakob Verbeek
- Contact: Maha Elbayad
- Publication: Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction
- URL: https://github.com/elbayadm/attn2d

6.6. Cyanure

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 FUNCTIONAL DESCRIPTION: version initiale

- Participant: Julien Mairal
- Contact: Julien Mairal
- URL: http://thoth.inrialpes.fr/people/mairal/arsenic/welcome.html

7. New Results

7.1. Visual Recognition and Robotics

7.1.1. Learning Disentangled Representations with Reference-Based Variational Autoencoders

Participants: Adria Ruiz, Oriol Martinez, Xavier Binefa, Jakob Verbeek.

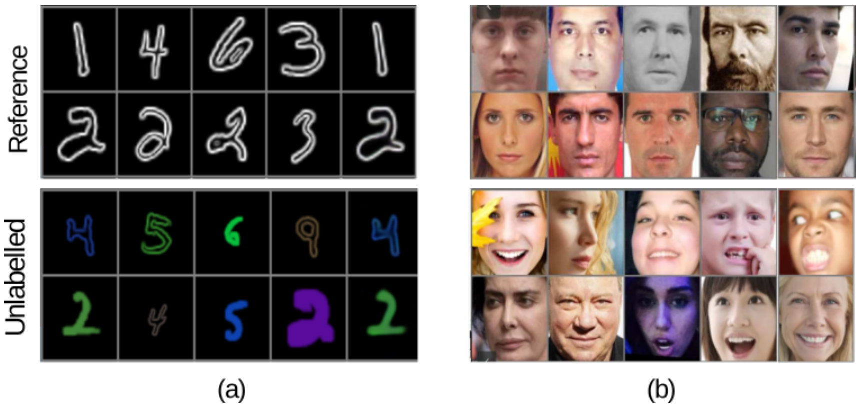


Figure 1. Illustration of different reference-based disentangling problems. (a) Disentangling style from digits. The reference distribution is composed by numbers with a fixed style (b) Disentangling factors of variations related with facial expressions. Reference images correspond to neutral faces. Note that pairing information between unlabelled and reference images is not available during training.

Learning disentangled representations from visual data, where different high-level generative factors are independently encoded, is of importance for many computer vision tasks. Supervised approaches, however, require a significant annotation effort in order to label the factors of interest in a training set. To alleviate the annotation cost, in [32] we introduce a learning setting which we refer to as “reference-based disentangling”. Given a pool of unlabelled images, the goal is to learn a representation where a set of target factors are disentangled from others. The only supervision comes from an auxiliary “reference set” that contains images where the factors of interest are constant. See Fig. 1 for illustrative examples. In order to address this problem, we propose reference-based variational autoencoders, a novel deep generative model designed to exploit the weak supervisory signal provided by the reference set. During training, we use the variational inference framework where adversarial learning is used to minimize the objective function. By addressing tasks such as feature learning, conditional image generation or attribute transfer, we validate the ability of the proposed model to learn disentangled representations from minimal supervision.

7.1.2. Tensor Decomposition and Non-linear Manifold Modeling for 3D Head Pose Estimation

Participants: Dmytro Derkach, Adria Ruiz, Federico M. Sukno.

Head pose estimation is a challenging computer vision problem with important applications in different scenarios such as human-computer interaction or face recognition. In [5], we present a 3D head pose estimation algorithm based on non-linear manifold learning. A key feature of the proposed approach is that it allows modeling the underlying 3D manifold that results from the combination of rotation angles. To do so, we use tensor decomposition to generate separate subspaces for each variation factor and show that each of them has a clear structure that can be modeled with cosine functions from a unique shared parameter per angle (see Fig. 2). Such representation provides a deep understanding of data behavior. We show that the proposed framework can be applied to a wide variety of input features and can be used for different purposes. Firstly, we test our system on a publicly available database, which consists of 2D images and we show that the cosine functions can be used to synthesize rotated versions from an object from which we see only a 2D image at a specific angle. Further, we perform 3D head pose estimation experiments using other two types of features: automatic landmarks and histogram-based 3D descriptors. We evaluate our approach on two publicly available databases, and demonstrate that angle estimations can be performed by optimizing the combination of these cosine functions to achieve state-of-the-art performance.

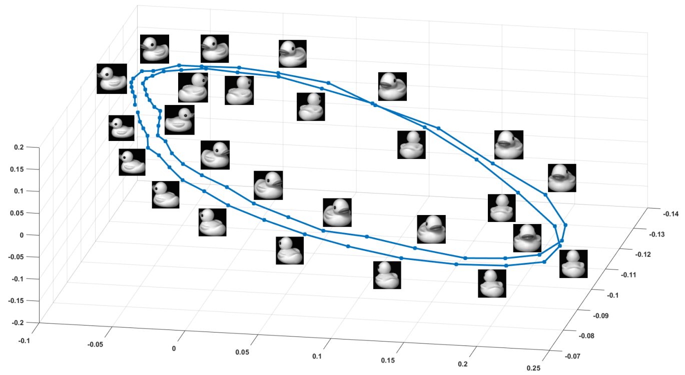


Figure 2. Visualization of the first three coefficients of the pose variation subspace for a dataset of single object rotated about the vertical axis.

7.1.3. Spreading vectors for similarity search

Participants: Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Hervé Jégou.

Discretizing multi-dimensional data distributions is a fundamental step of modern indexing methods. State-of-the-art techniques learn parameters of quantizers on training data for optimal performance, thus adapting quantizers to the data. In this work [29], we propose to reverse this paradigm and adapt the data to the quantizer: we train a neural net which last layer forms a fixed parameter-free quantizer, such as pre-defined points of a hyper-sphere. As a proxy objective, we design and train a neural network that favors uniformity in the spherical latent space, while preserving the neighborhood structure after the mapping. We propose a new regularizer derived from the Kozachenko–Leonenko differential entropy estimator to enforce uniformity and combine it with a locality-aware triplet loss. Experiments show that our end-to-end approach outperforms most learned quantization methods, and is competitive with the state of the art on widely adopted benchmarks. Furthermore, we show that training without the quantization step results in almost no difference in accuracy, but yields a generic catalyzer 3 that can be applied with any subsequent quantizer. The code is available online.

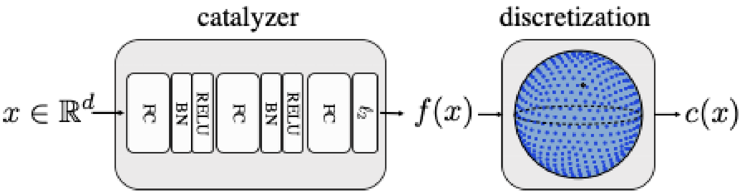


Figure 3. Our method learns a network that encodes the input space $\mathbb{R}^d$ into a code $c(x)$. It is learned end-to-end, yet the part of the network in charge of the discretization operation is fixed in advance, thereby avoiding optimization problems. The learnable function f, namely the “catalyzer", is optimized to increase the quality of the subsequent coding stage.

7.1.4. Diversity with Cooperation: Ensemble Methods for Few-Shot Classification

Participants: Nikita Dvornik, Cordelia Schmid, Julien Mairal.

Few-shot classification consists of learning a predictive model that is able to effectively adapt to a new class, given only a few annotated samples. To solve this challenging problem, meta-learning has become a popular paradigm that advocates the ability to “learn to adapt". Recent works have shown, however, that simple learning strategies without meta-learning could be competitive. In our ICCV’19 paper [17], we go a step further and show that by addressing the fundamental high-variance issue of few-shot learning classifiers, it is possible to significantly outperform current meta-learning techniques. Our approach consists of designing an ensemble of deep networks to leverage the variance of the classifiers, and introducing new strategies to encourage the networks to cooperate, while encouraging prediction diversity, as illustrated in Figure 4. Evaluation is conducted on the mini-ImageNet and CUB datasets, where we show that even a single network obtained by distillation yields state-of-the-art results.

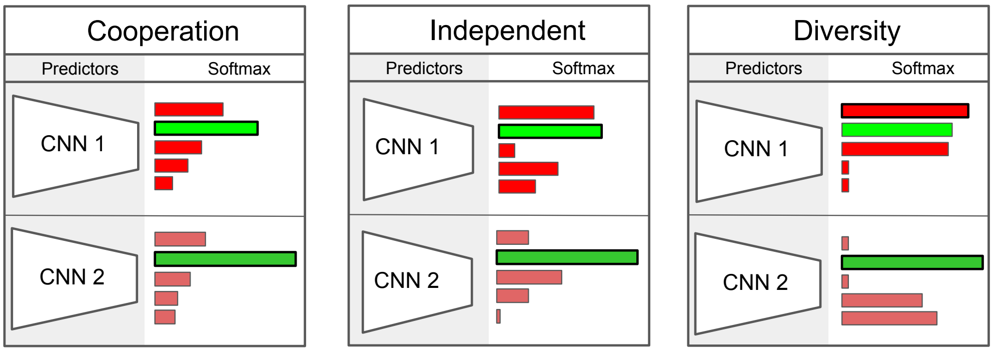


*Figure 4. **Illustration of the cooperation and diversity strategies on two networks.** All networks receive the same image as input and compute corresponding class probabilities with softmax. Cooperation encourages the non-ground truth probabilities (in red) to be similar, after normalization, whereas diversity encourages orthogonality.*

7.1.5. Unsupervised Pre-Training of Image Features on Non-Curated Data

Participants: Mathilde Caron, Piotr Bojanowski [Facebook AI], Julien Mairal, Armand Joulin [Facebook AI].

Pre-training general-purpose visual features with convolutional neural networks without relying on annotations is a challenging and important task. Most recent efforts in unsupervised feature learning have focused on either small or highly curated datasets like ImageNet, whereas using non-curated raw datasets was found to decrease the feature quality when evaluated on a transfer task. Our goal is to bridge the performance gap between unsupervised methods trained on curated data, which are costly to obtain, and massive raw datasets that are easily available. To that effect, we propose a new unsupervised approach, DeeperCluster [13], described in Figure 5 which leverages self-supervision and clustering to capture complementary statistics from large-scale data. We validate our approach on 96 million images from YFCC100M, achieving state-of-the-art results among unsupervised methods on standard benchmarks, which confirms the potential of unsupervised learning when only non-curated raw data are available. We also show that pre-training a supervised VGG-16 with our method achieves 74.9% top-1 classification accuracy on the validation set of ImageNet, which is an improvement of +0.8% over the same network trained from scratch.

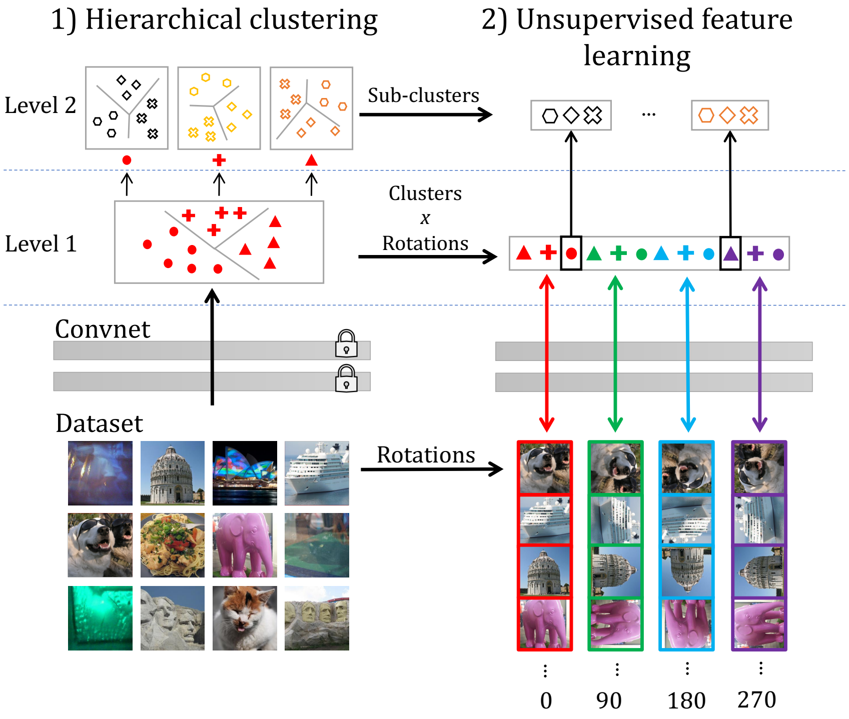


Figure 5. DeeperCluster alternates between a hierachical clustering of the features and learning the parameters of a convnet by predicting both the rotation angle and the cluster assignments in a single hierachical loss.

7.1.6. Learning to Augment Synthetic Images for Sim2Real Policy Transfer

Participants: Alexander Pashevich, Robin Strudel [Inria WILLOW], Igor Kalevatykh [Inria WILLOW], Ivan Laptev [Inria WILLOW], Cordelia Schmid.

Vision and learning have made significant progress that could improve robotics policies for complex tasks and environments. Learning deep neural networks for image understanding, however, requires large amounts of domain-specific visual data. While collecting such data from real robots is possible, such an approach limits the scalability as learning policies typically requires thousands of trials. In this work [25] we attempt to learn manipulation policies in simulated environments. Simulators enable scalability and provide access to the underlying world state during training. Policies learned in simulators, however, do not transfer well to real scenes given the domain gap between real and synthetic data. We follow recent work on domain randomization and augment synthetic images with sequences of random transformations. Our main contribution is to optimize the augmentation strategy for sim2real transfer and to enable domain-independent policy learning, as illustrated in Figure 6. We design an efficient search for depth image augmentations using object localization as a proxy task. Given the resulting sequence of random transformations, we use it to augment synthetic depth images during policy learning. Our augmentation strategy is policy-independent and enables policy learning with no real images. We demonstrate our approach to significantly improve accuracy on three manipulation tasks evaluated on a real robot.

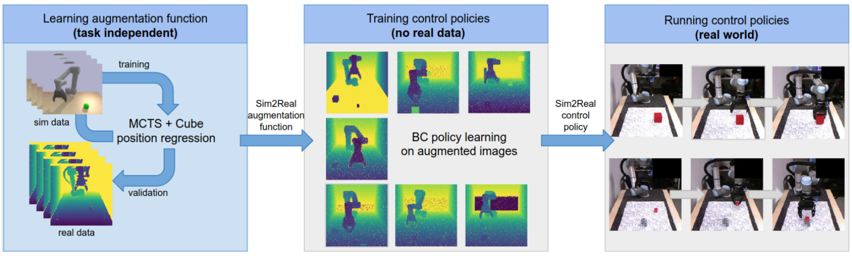

Learning augmentation function
(task independent)
training
sim data
MCTS + Cube position regression
validation
real data
Sim2Real augmentation function
Training control policies
(no real data)
BC policy learning on augmented images
Sim2Real control policy
Running control policies
(real world)

Figure 6. Overview of the method. Our contribution is the policy-independent learning of depth image augmentations (left). The resulting sequence of augmentations is applied to synthetic depth images while learning manipulation policies in a simulator (middle). The learned policies are directly applied to real robot scenes without finetuning on real images.

7.1.7. Learning to combine primitive skills: A step towards versatile robotic manipulation

Participants: Robin Strudel [Inria WILLOW], Alexander Pashevich, Igor Kalevatykh [Inria WILLOW], Ivan Laptev [Inria WILLOW], Josef Sivic [Inria WILLOW], Cordelia Schmid.

Manipulation tasks such as preparing a meal or assembling furniture remain highly challenging for robotics and vision. Traditional task and motion planning (TAMP) methods can solve complex tasks but require full state observability and are not adapted to dynamic scene changes. Recent learning methods can operate directly on visual inputs but typically require many demonstrations and/or task-specific reward engineering. In this work [40] we aim to overcome previous limitations and propose a reinforcement learning (RL) approach to task planning that learns to combine primitive skills illustrated in Figure 7. First, compared to previous learning methods, our approach requires neither intermediate rewards nor complete task demonstrations during training. Second, we demonstrate the versatility of our vision-based task planning in challenging settings with temporary occlusions and dynamic scene changes. Third, we propose an efficient training of basic skills from few synthetic demonstrations by exploring recent CNN architectures and data augmentation. Notably, while all of our policies are learned on visual inputs in simulated environments, we demonstrate the successful transfer and high success rates when applying such policies to manipulation tasks on a real UR5 robotic arm.

7.1.8. Probabilistic Reconstruction Networks for 3D Shape Inference from a Single Image

Participants: Roman Klokov, Jakob Verbeek, Edmond Boyer [Inria Morpheo].

In our BMVC'19 paper [21], we study end-to-end learning strategies for 3D shape inference from images, in particular from a single image. Several approaches in this direction have been investigated that explore different shape representations and suitable learning architectures. We focus instead on the underlying probabilistic mechanisms involved and contribute a more principled probabilistic inference-based reconstruction framework, which we coin Probabilistic Reconstruction Networks. This framework expresses image conditioned

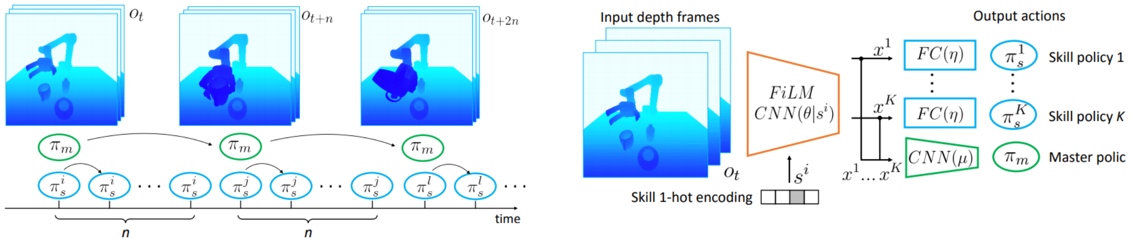


Figure 7. Illustration of our approach. (Left): Temporal hierarchy of master and skill policies. The master policy π_m is executed at a coarse interval of n time-steps to select among K skill policies $\pi_s^1 \ldots \pi_s^K$. Each skill policy generates control for a primitive action such as grasping or pouring. (Right): CNN architecture used for the skill and master policies.

3D shape inference through a family of latent variable models, and naturally decouples the choice of shape representations from the inference itself. Moreover, it suggests different options for the image conditioning and allows training in two regimes, using either Monte Carlo or variational approximation of the marginal likelihood. Using our Probabilistic Reconstruction Networks we obtain single image 3D reconstruction results that set a new state of the art on the ShapeNet dataset in terms of the intersection over union and earth mover's distance evaluation metrics. Interestingly, we obtain these results using a basic voxel grid representation, improving over recent work based on finer point cloud or mesh based representations. In Figure 8 we show a schematic overview of our model.

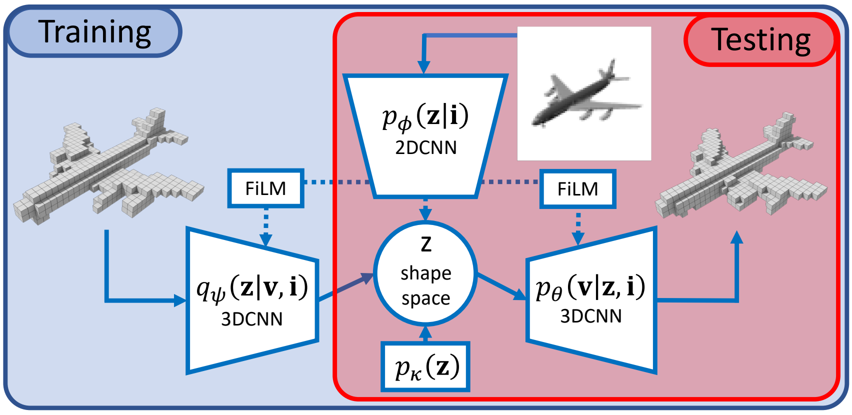

Training
Testing
$p_\phi(\mathbf{z}|\mathbf{i})$
2DCNN
FiLM
FiLM
$q_\psi(\mathbf{z}|\mathbf{v},\mathbf{i})$
3DCNN
z
shape
space
$p_\theta(\mathbf{v}|\mathbf{z},\mathbf{i})$
3DCNN
$p_\kappa(\mathbf{z})$

Figure 8. Probabilistic Reconstruction Networks for 3D shape inference from a single image. Arrows show the computational flow through the model, dotted arrows show optional image conditioning. Conditioning between 2D and 3D tensors is achieved by means of FiLM layers. The inference network q_ψ is only used during training for variational inference.

7.1.9. Hierarchical Scene Coordinate Classification and Regression for Visual Localization

Participants: Xiaotian Li [Aalto Univ., Finland], Shuzhe Wang [Aalto Univ., Finland], Li Zhao [Aalto Univ., Finland], Jakob Verbeek, Juho Kannala [Aalto Univ., Finland].

Visual localization is critical to many applications in computer vision and robotics. To address single-image RGB localization, state-of-the-art feature-based methods match local descriptors between a query image and a pre-built 3D model. Recently, deep neural networks have been exploited to regress the mapping between raw pixels and 3D coordinates in the scene, and thus the matching is implicitly performed by the forward pass through the network. However, in a large and ambiguous environment, learning such a regression task directly can be difficult for a single network. In our paper [37], we present a new hierarchical scene coordinate network to predict pixel scene coordinates in a coarse-to-fine manner from a single RGB image. The network consists of a series of output layers with each of them conditioned on the previous ones. The final output layer predicts the 3D coordinates and the others produce progressively finer discrete location labels. The proposed method outperforms the baseline regression-only network and allows us to train single compact models which scale robustly to large environments. It sets a new state-of-the-art for single-image RGB localization performance on the 7-Scenes, 12-Scenes, Cambridge Landmarks datasets, and three combined scenes. Moreover, for large-scale outdoor localization on the Aachen Day-Night dataset, our approach is much more accurate than existing scene coordinate regression approaches, and reduces significantly the performance gap w.r.t. explicit feature matching approaches. In Figure 9 we illustrate the scene coordinate predictions for the Aachen dataset experiments.

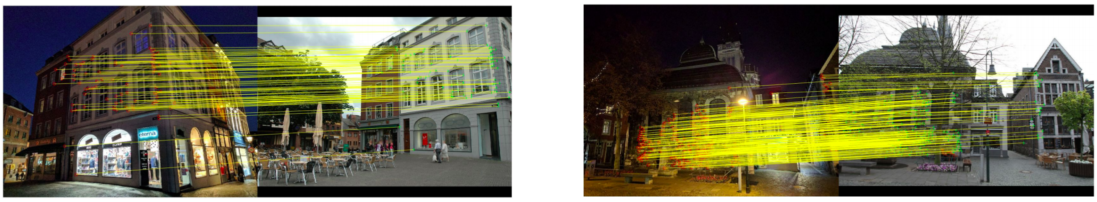


Figure 9. The scene coordinate predictions are visualized as 2D-2D matches between the query (left) and database (right) images. For each pair, the retrieved database image with the largest number of inliers is selected, and only the inlier matches are visualized. We show that our method is able to produce accurate correspondences for challenging queries.

7.1.10. Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images

Participants: Valentin Gabeur, Jean-Sébastien Franco [Inria Morpheo], Xavier Martin, Cordelia Schmid, Gregory Rogez [NAVER LABS Europe].

While the recent progress in convolutional neural networks has allowed impressive results for 3D human pose estimation, estimating the full 3D shape of a person is still an open issue. Model-based approaches can output precise meshes of naked under-cloth human bodies but fail to estimate details and un-modelled elements such as hair or clothing. On the other hand, non-parametric volumetric approaches can potentially estimate complete shapes but, in practice, they are limited by the resolution of the output grid and cannot produce detailed estimates. In this paper [19], we propose a non-parametric approach that employs a double depth map 10 to represent the 3D shape of a person: a visible depth map and a “hidden” depth map are estimated and combined, to reconstruct the human 3D shape as done with a “mould”. This representation through 2D depth maps allows a higher resolution output with a much lower dimension than voxel-based volumetric representations.

7.1.11. Focused Attention for Action Recognition

Participants: Vladyslav Sydorov, Karteek Alahari.

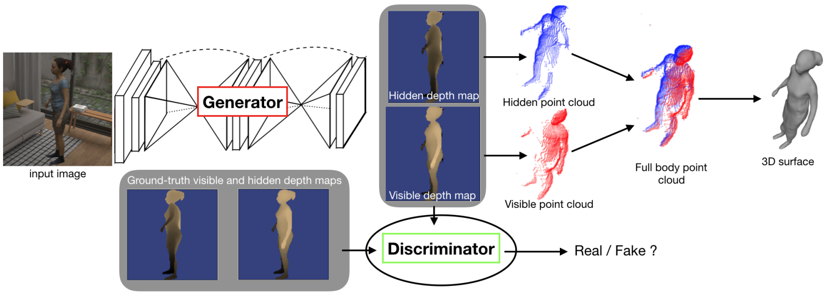


Figure 10. Given a single image, we estimate the "visible" and the "hidden" depth maps. The 3D point clouds of these 2 depth maps are combined to form a full-body 3D point cloud, as if lining up the 2 halves of a "mould". The 3D shape is then reconstructed using Poisson reconstruction. An adversarial training with a discriminator is employed to increase the humanness of the estimation.

In this paper [30], we introduce an attention model for video action recognition that allows processing video in higher resolution, by focusing on the relevant regions first. The network-specific saliency is utilized to guide the cropping, we illustrate the procedure in Figure 11. We show performance improvement on the Charades dataset with this strategy.

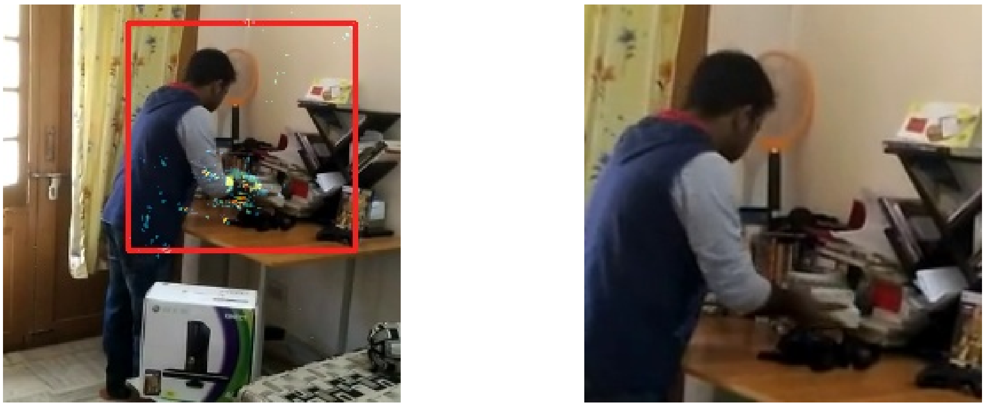


Figure 11. Example of attention on Charades action recognition dataset. (Left) Saliency scores (displayed as a heatmap) are localized around the object, a box maximizing the saliency measure within is selected. (Right) The network is provided with the relevant crop of the video, and can process it at a higher resolution.

7.2. Statistical Machine Learning

7.2.1. 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 , 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.

7.2.2. A Generic Acceleration Framework for Stochastic Composite Optimization

Participants: Andrei Kulunchakov, Julien Mairal.

In [35], we introduce various mechanisms to obtain accelerated first-order stochasticoptimization algorithms when the objective function is convex or stronglyconvex. Specifically, we extend the Catalyst approach originally designed fordeterministic objectives to the stochastic setting. Given an optimizationmethod with mild convergence guarantees for strongly convex problems,the challenge is to accelerate convergence to a noise-dominated region, andthen achieve convergence with an optimal worst-case complexity depending on thenoise variance of the gradients.A side contribution of our work is also a generic analysis that canhandle inexact proximal operators, providing new insights about the robustness of stochastic algorithms when the proximal operator cannot be exactly computed. An illustration from this work is explained in Figure 12.

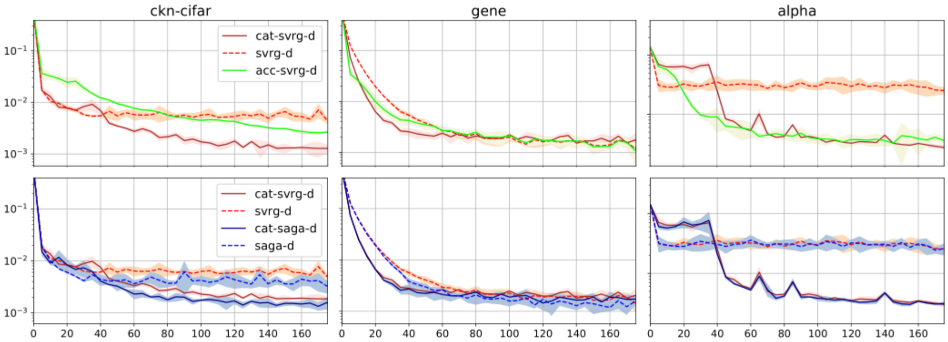


Figure 12. Accelerating SVRG-like (top) and SAGA (bottom) methods for ℓ_2-logistic regression with $\mu = 1/(100n)$ (bottom) for mild dropout, which imitates stochasticity in the gradients. All plots are on a logarithmic scale for the objective function value, and the x-axis denotes the number of epochs. The colored tubes around each curve denote a standard deviations across 5 runs. The curves show that acceleration may be useful even in the stochastic optimization regime.

7.2.3. Estimate Sequences for Variance-Reduced Stochastic Composite Optimization

Participants: Andrei Kulunchakov, Julien Mairal.

In [23], we propose a unified view of gradient-based algorithms for stochastic convex composite optimization. By extending the concept of estimate sequence introduced by Nesterov, we interpret a large class of stochastic optimization methods as procedures that iteratively minimize a surrogate of the objective. This point of view covers stochastic gradient descent (SGD), the variance-reduction approaches SAGA, SVRG, MISO, their proximal variants, and has several advantages: (i) we provide a simple generic proof of convergence for all of the aforementioned methods; (ii) we naturally obtain new algorithms with the same guarantees; (iii) we derive generic strategies to make these algorithms robust to stochastic noise, which is useful when data is corrupted by small random perturbations. Finally, we show that this viewpoint is useful to obtain accelerated algorithms. A comparison with different approaches is shown in Figure 13.

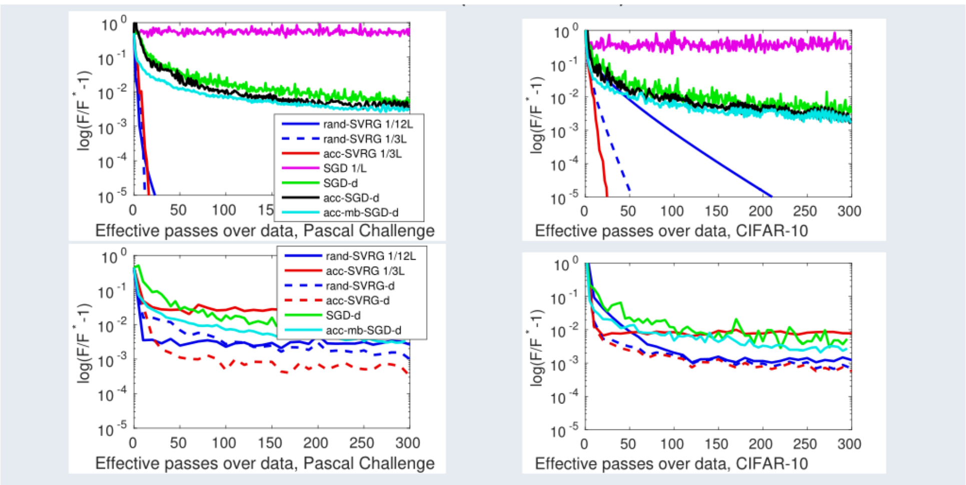


Figure 13. Comparison of different standard approaches with our developed method on two datasets for ℓ_2-logistic regression with mild dropout (bottom) and deterministic case (above). The case of exact gradient computations clearly shows benefits from acceleration, which consist in fast linear convergence. In the stochastic case, we demonstrate either superiority or high competitiveness of the developed method along with its unbiased convergence to the optimum. In both cases, we show that acceleration is able to generically comprise strengths of standard methods and even outperform them.

7.2.4. White-box vs Black-box: Bayes Optimal Strategies for Membership Inference

Participants: Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, Hervé Jégou.

Membership inference determines, given a sample and trained parameters of a machine learning model, whether the sample was part of the training set. In this paper [28], we derive the optimal strategy for membership inference with a few assumptions on the distribution of the parameters. We show that optimal attacks only depend on the loss function, and thus black-box attacks are as good as white-box attacks. As the optimal strategy is not tractable, we provide approximations of it leading to several inference methods 14, and show that existing membership inference methods are coarser approximations of this optimal strategy. Our membership attacks outperform the state of the art in various settings, ranging from a simple logistic regression to more complex architectures and datasets, such as ResNet-101 and Imagenet.

7.3. Theory and Methods for Deep Neural Networks

7.3.1. Group Invariance, Stability to Deformations, and Complexity of Deep Convolutional Representations

Participants: Alberto Bietti, Julien Mairal.

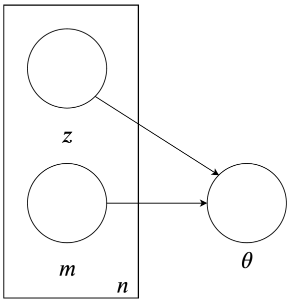


Figure 14. Plate notation of the membership inference problem: for each data point z_i, a binary membership variable m_i is sampled, and z_i belongs to the training set iff $m_i = 1$. Given the trained parameters θ and a sample z_i, we want to infer the value of m_i.

The success of deep convolutional architectures is often attributed in part to their ability to learn multiscale and invariant representations of natural signals. However, a precise study of these properties and how they affect learning guarantees is still missing. In the paper [3], we consider deep convolutional representations of signals; we study their invariance to translations and to more general groups of transformations, their stability to the action of diffeomorphisms, and their ability to preserve signal information. This analysis is carried by introducing a multilayer kernel based on convolutional kernel networks and by studying the geometry induced by the kernel mapping. We then characterize the corresponding reproducing kernel Hilbert space (RKHS), showing that it contains a large class of convolutional neural networks with homogeneous activation functions. This analysis allows us to separate data representation from learning, and to provide a canonical measure of model complexity, the RKHS norm, which controls both stability and generalization of any learned model. In addition to models in the constructed RKHS, our stability analysis also applies to convolutional networks with generic activations such as rectified linear units, and we discuss its relationship with recent generalization bounds based on spectral norms.

7.3.2. A Kernel Perspective for Regularizing Deep Neural Networks

Participants: Alberto Bietti, Grégoire Mialon, Dexiong Chen, Julien Mairal.

We propose a new point of view for regularizing deep neural networks by using the norm of a reproducing kernel Hilbert space (RKHS) [12]. Even though this norm cannot be computed, it admits upper and lower approximations leading to various practical strategies. Specifically, this perspective (i) provides a common umbrella for many existing regularization principles, including spectral norm and gradient penalties, or adversarial training, (ii) leads to new effective regularization penalties, and (iii) suggests hybrid strategies combining lower and upper bounds to get better approximations of the RKHS norm. We experimentally show this approach to be effective when learning on small datasets, or to obtain adversarially robust models.

7.3.3. On the Inductive Bias of Neural Tangent Kernels

Participants: Alberto Bietti, Julien Mairal.

State-of-the-art neural networks are heavily over-parameterized, making the optimization algorithm a crucial ingredient for learning predictive models with good generalization properties. A recent line of work has shown that in a certain over-parameterized regime, the learning dynamics of gradient descent are governed by a certain kernel obtained at initialization, called the neural tangent kernel. In [12], we study the inductive bias of learning in such a regime by analyzing this kernel and the corresponding function space (RKHS). In particular, we study smoothness, approximation, and stability properties of functions with finite norm, including stability to image deformations in the case of convolutional networks, and compare to other known kernels for similar architectures.

7.3.4. *Large Memory Layers with Product Keys*

Participants: Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou.

This paper introduces a structured memory which can be easily integrated into a neural network. The memory is very large by design and significantly increases the capacity of the architecture, by up to a billion parameters with a negligible computational overhead. Its design and access pattern is based on product keys, which enable fast and exact nearest neighbor search. The ability to increase the number of parameters while keeping the same computational budget lets the overall system strike a better trade-off between prediction accuracy and computation efficiency both at training and test time. This memory layer, shown in Figure 15, allows us to tackle very large scale language modeling tasks. In our experiments we consider a dataset with up to 30 billion words, and we plug our memory layer in a state-of-the-art transformer-based architecture. In particular, we found that a memory augmented model with only 12 layers outperforms a baseline transformer model with 24 layers, while being twice faster at inference time. We release our code for reproducibility purposes.

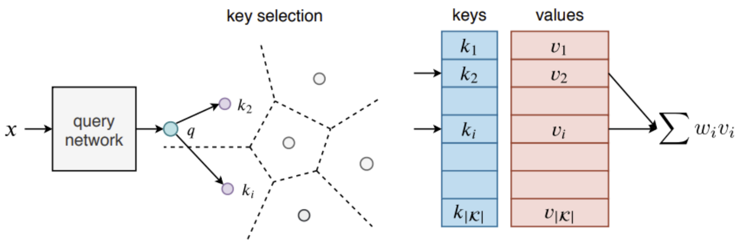


Figure 15. Overview of a key-value memory layer: The input x is processed through a query network that produces a query vector q, which is compared to all the keys. The output is the sparse weighted sum over the memories associated with the selected keys. For a large number of keys $|\mathcal{K}|$, the key selection procedure becomes too expensive in practice. Our product key method is exact and makes this search process very fast.

7.3.5. *Understanding Priors in Bayesian Neural Networks at the Unit Level*

Participants: Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo [Univ. Granada, Spain], Julyan Arbel [Inria MISTIS].

In our ICML'19 paper [31], we investigate deep Bayesian neural networks with Gaussian weight priors and a class of ReLUlike nonlinearities. Bayesian neural networks with Gaussian priors are well known to induce an L2, "weight decay", regularization. Our results characterize a more intricate regularization effect at the level of the unit activations. Our main result establishes that the induced prior distribution on the units before and after activation becomes increasingly heavy-tailed with the depth of the layer. We show that first layer units are Gaussian, second layer units are sub-exponential, and units in deeper layers are characterized by sub-Weibull distributions. Our results provide new theoretical insight on deep Bayesian neural networks, which we corroborate with experimental simulation results.

7.3.6. *Adaptative Inference Cost With Convolutional Neural Mixture Models*

Participants: Adria Ruiz, Jakob Verbeek.

Despite the outstanding performance of convolutional neural networks (CNNs) for many vision tasks, the required computational cost during inference is problematic when resources are limited. In this paper [27], we propose Convolutional Neural Mixture Models (CNMMs), a probabilistic model embedding a large number of CNNs that can be jointly trained and evaluated in an efficient manner. Within the proposed framework, we present different mechanisms to prune subsets of CNNs from the mixture, allowing to easily adapt the computational cost required for inference (see Fig. 16). Image classification and semantic segmentation experiments show that our method achieve excellent accuracy-compute trade-offs. Moreover, unlike most of previous approaches, a single CNMM provides a large range of operating points along this trade-off, without any re-training.

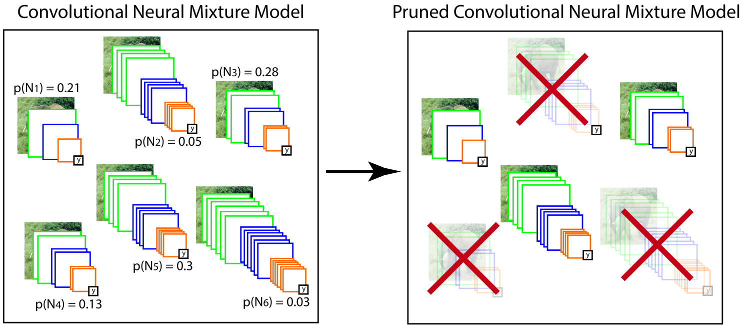


Figure 16. A Convolutional Neural Mixture Model embeds a large number of CNNs. Weight sharing enables efficient joint training of all networks and computation of the mixture output. The learned mixing weights can be used to remove networks from the mixture, and thus reduce the computational cost of inference.

7.4. Pluri-disciplinary Research

7.4.1. Biological Sequence Modeling with Convolutional Kernel Networks

Participants: Dexiong Chen, Laurent Jacob, Julien Mairal.

The growing number of annotated biological sequences available makes it possible to learn genotype-phenotype relationships from data with increasingly high accuracy. When large quan- tities of labeled samples are available for training a model, convolutional neural networks can be used to predict the phenotype of unannotated sequences with good accuracy. Unfortunately, their performance with medium- or small-scale datasets is mitigated, which requires inventing new data-efficient approaches. In this paper [4], [14], we introduce a hybrid approach between convolutional neural networks and kernel methods to model biological sequences. Our method, shown in Figure 17, enjoys the ability of convolutional neural networks to learn data representations that are adapted to a specific task, while the kernel point of view yields algorithms that perform significantly better when the amount of training data is small. We illustrate these advantages for transcription factor binding prediction and protein homology detection, and we demonstrate that our model is also simple to interpret, which is crucial for discovering predictive motifs in sequences. The source code is freely available at https://gitlab.inria.fr/dchen/CKN-seq.

7.4.2. Recurrent Kernel Networks

Participants: Dexiong Chen, Laurent Jacob [CNRS, LBBE Laboratory], Julien Mairal.

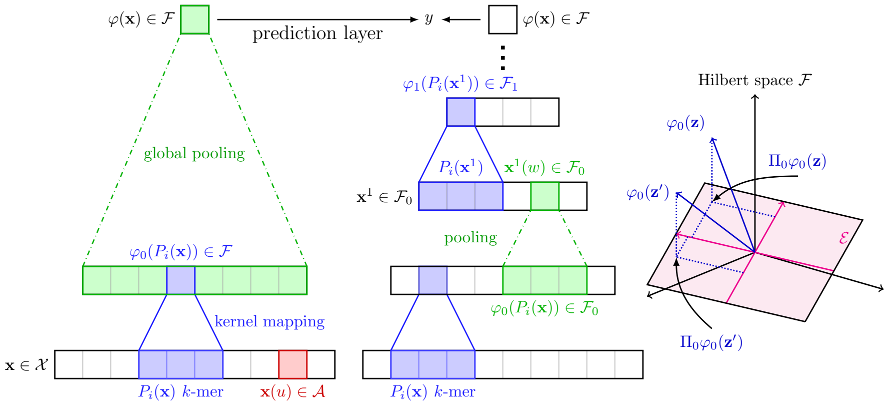

$\varphi(\mathbf{x}) \in \mathcal{F}$
prediction layer
y
$\varphi(\mathbf{x}) \in \mathcal{F}$
$\varphi_1(P_i(\mathbf{x}^1)) \in \mathcal{F}_1$
Hilbert space $\mathcal{F}$
$\varphi_0(\mathbf{z})$
$\Pi_0\varphi_0(\mathbf{z})$
$\varphi_0(\mathbf{z}')$
$\mathcal{E}$
$\Pi_0\varphi_0(\mathbf{z}')$
global pooling
$P_i(\mathbf{x}^1)$
$\mathbf{x}^1(w) \in \mathcal{F}_0$
$\mathbf{x}^1 \in \mathcal{F}_0$
pooling
$\varphi_0(P_i(\mathbf{x})) \in \mathcal{F}$
$\varphi_0(P_i(\mathbf{x})) \in \mathcal{F}_0$
kernel mapping
$\mathbf{x} \in \mathcal{X}$
$P_i(\mathbf{x})$ k-mer
$\mathbf{x}(u) \in \mathcal{A}$
$P_i(\mathbf{x})$ k-mer

Figure 17. Construction of single-layer (left) and multilayer (middle) CKN-seq and the approximation of one layer (right). For a single-layer model, each k-mer $P_i(\mathbf{x})$ is mapped to $\varphi_0(P_i(\mathbf{x}))$ in $\mathcal{F}$ and projected to $\Pi\varphi_0(P_i(\mathbf{x}))$ parametrized by $\psi_0(P_i(\mathbf{x}))$. Then, the final finite-dimensional sequence is obtained by the global pooling, $\psi(\mathbf{x}) = \frac{1}{m}\sum_{i=0}^{m}\psi_0(P_i(\mathbf{x}))$. The multilayer construction is similar, but relies on intermediate maps, obtained by local pooling.

Substring kernels are classical tools for representing biological sequences or text. However, when large amounts of annotated data are available, models that allow end-to-end training such as neural networks are often preferred. Links between recurrent neural networks (RNNs) and substring kernels have recently been drawn, by formally showing that RNNs with specific activation functions were points in a reproducing kernel Hilbert space (RKHS). In this paper [15], we revisit this link by generalizing convolutional kernel networks—–originally related to a relaxation of the mismatch kernel—–to model gaps in sequences. It results in a new type of recurrent neural network (Figure 18), which can be trained end-to-end with backpropagation, or without supervision by using kernel approximation techniques. We experimentally show that our approach is well suited to biological sequences, where it outperforms existing methods for protein classification tasks.

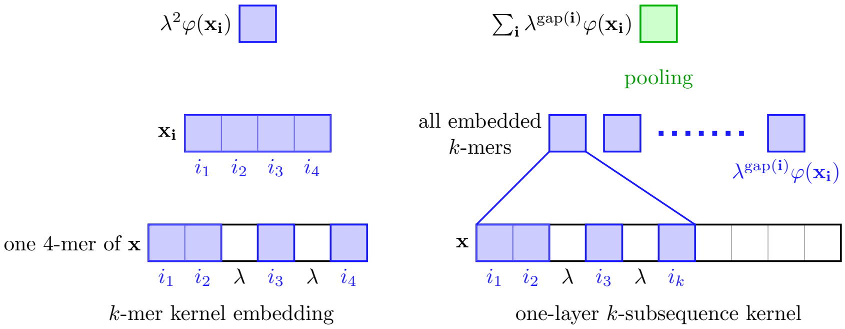
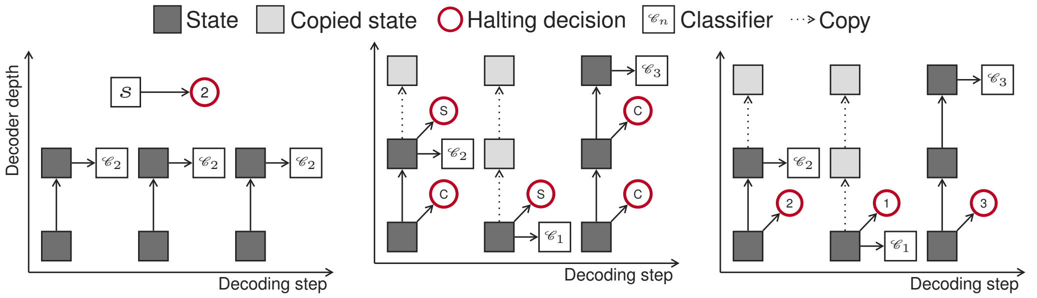

$\lambda^2\varphi(\mathbf{x_i})$
$\sum_\mathbf{i} \lambda^{\mathrm{gap}(\mathbf{i})}\varphi(\mathbf{x_i})$
pooling
$\mathbf{x_i}$
all embedded k-mers
i_1 i_2 i_3 i_4
$\lambda^{\mathrm{gap}(\mathbf{i})}\varphi(\mathbf{x_i})$
one 4-mer of $\mathbf{x}$
$\mathbf{x}$
i_1 i_2 λ i_3 λ i_4
i_1 i_2 λ i_3 λ i_k
k-mer kernel embedding
one-layer k-subsequence kernel

Figure 18. Representation of a sequence in a RKHS based on our kernel.

7.4.3. Depth-adaptive Transformer

Participants: Maha Elbayad, Jiatao Gu [Facebook AI], Edouard Grave [Facebook AI], Michael Auli [Facebook AI].

State of the art sequence-to-sequence models for large scale tasks perform a fixed number of computations for each input sequence regardless of whether it is easy or hard to process. In our ICLR'2020 paper [18], we train Transformer models which can make output predictions at different stages of the network and we investigate different ways to predict how much computation is required for a particular sequence. Unlike dynamic computation in Universal Transformers, which applies the same set of layers iteratively, we apply different layers at every step to adjust both the amount of computation as well as the model capacity. On IWSLT German-English translation our approach matches the accuracy of a well tuned baseline Transformer while using less than a quarter of the decoder layers. Figure 19 illustrates the different halting mechanisms investigated in this work. Namely, a sequence-level approach where we assume all the sequence's tokens are equally difficult and a token-level approach where tokens exit at varying depths.



(a) Sequence-level depth conditioned on the encoding of the source sequence. We exit at the same depth for all tokens in the sequence.

(b) Token-level depth with a geometric-like distribution. At each gate we decide to continue (C) through the decoder's blocks or to stop (S) and emit an output.

(c) Token-level depth with a multinomial distribution. After the first decoder block, we predict the appropriate exit for the current step (in this example from 1 to 3).

Figure 19. Illustration of the variant adaptive depth predictors: (a) the sequence-level and (b, c) at the token-level.

8. Bilateral Contracts and Grants with Industry

8.1. Intel

Participants: Cordelia Schmid, Karteek Alahari.

The Intel Network on Intelligent Systems in Europe brings together leading researchers in robotics, computer vision, motor control, and machine learning. We are part of this network and have participated in the annual retreat in 2018. Funding will be provided on an annual basis, every year, as long as we are part of the network.

8.2. Facebook

Participants: Cordelia Schmid, Jakob Verbeek, Julien Mairal, Karteek Alahari, Pauline Luc, Alexandre Sablayrolles, Mathilde Caron, Lina Mezghani.

The collaboration started in 2016. The topics include image retrieval with CNN based descriptors, weakly supervised object detection and semantic segmentation, and learning structured models for action recognition in videos. In 2016, Pauline Luc started her PhD funded by a CIFRE grant, jointly supervised by Jakob Verbeek (Inria) and Camille Couprie (Facebook AI Research). THOTH has been selected in 2016 as a recipient for the Facebook GPU Partnership program. In this context Facebook has donated two state-of-the-art servers with 8 GPUs. In 2017, Alexandre Sablayrolles started his CIFRE grant, jointly supervised by Cordelia Schmid, and Herve Jegou and Matthijs Douze at Facebook AI Research. In 2018, Mathilde Caron started as a CIFRE PhD student, jointly supervised by Julien Mairal, and Armand Joulin and Piotr Bojanowski at Facebook AI Research. Lina Mezghani is the new PhD student in this collaboration since 2019.

8.3. NAVER LABS Europe

Participant: Karteek Alahari.

This collaboration started when NAVER LABS Europe was Xerox Research Centre Europe, and has been on-going since October 2009 with two co-supervised CIFRE scholarships (2009–2012, 2011-2014). Starting June 2014 we signed a third collaborative agreement for a duration of three years. The goal is to develop approaches for deep learning based image description and pose estimation in videos. Jakob Verbeek and Diane Larlus (XRCE) jointly supervise a PhD-level intern for a period of 6 months in 2016-2017. XRCE then became Naver in 2017. A one-year research contract on action recognition in videos started in Sep 2017. The approach developed by Vasileios Choutas implements pose-based motion features, which are shown to be complementary to state-of-the-art I3D features. Nieves Crasto's internship in 2018 was jointly supervised by Philippe Weinzaepfel (NAVER LABS), Karteek Alahari and Cordelia Schmid. A new CIFRE PhD contract was submitted to ANRT for approval in October 2019.

8.4. Valeo AI

Participants: Karteek Alahari, Florent Bartoccioni.

This collaboration started in 2019 with the arrival of PhD student Florent Bartoccioni. Despite the progress seen in computer vision, artificial systems lack the capability to address the large disparity between human and machine-based scene understanding. For example, at any road intersection most people have the ability to accurately forecast or anticipate events in this scenario, such as changes in colour of the traffic lights, when and how pedestrians are likely to cross the street. This apparently natural human behaviour is not replicable by state-of-theart computer vision methods, which are ill-equipped to make such forecasts. The goal of this collaborative PhD is to address this forecasting problem.

8.5. Criteo

Participant: Julien Mairal.

This collaboration started in April 2019, with the arrival of a master student, Houssam Zenati, who will pursue a CIFRE PhD starting in 2020. The goal of this collaboration is to develop machine learning techniques for counterfactual loss optimization, which is a fundamental problem in machine learning related to causal inference. The goal is to learn stochastic policies, based on offline logged data. The problem is important for web advertising, which is the main activity of the Criteo company, but the potential scope of application is much larger, with possible applications in medicine and experimental sciences.

8.6. Google

Participants: Karteek Alahari, Minttu Alakuijala, Valentin Gabeur, Julien Mairal.

This collaboration started in February 2019, with the arrival of two CIFRE PhD students, Minttu Alakuijala and Valentin Gabeur, who are respectively working on visual models for robotics, and 3D human pose estimation.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. MIAI chair - Towards more data efficiency in machine learning

Participants: Julien Mairal, Karteek Alahari, Jakob Verbeek.

Julien Mairal holds a chair of the 3IA MIAI institute. The goal is to improve the data efficiency of machine learning algorithms.

9.1.2. MIAI chair - Towards self-supervised visual learning

Participant: Cordelia Schmid.

Cordelia Schmid holds a chair of the 3IA MIAI institute. The goal is to develop new self-supervised learning methods for computer vision.

9.1.3. MIAI chair - Multiscale, multimodal and multitemporal remote sensing

Participant: Jocelyn Chanussot.

Jocelyn Chanussot holds a chair of the 3IA MIAI institute.

9.1.4. DeCore (Deep Convolutional and Recurrent networks for image, speech, and text)

Participants: Jakob Verbeek, Maha Elbayad.

DeCore is a project-team funded by the Persyval Lab for 3.5 years (september 2016 - February 2020), coordinated by Jakob Verbeek. It unites experts from Grenoble's applied-math and computer science labs LJK, GIPSA-LAB and LIG in the areas of computer vision, machine learning, speech, natural language processing, and information retrieval. The purpose of DeCore is to stimulate collaborative interdisciplinary research on deep learning in the Grenoble area, which is likely to underpin future advances in machine perception (vision, speech, text) over the next decade. It provides funding for two full PhD students. Maha Elbayad is one of them, supervised by Jakob Verbeek and Laurant Besacier (LIG, UGA).

9.1.5. PEPS AMIES AuMalis POLLEN

Participant: Karteek Alahari.

This is a collaborative project with POLLEN, a startup in the Grenoble area, which develops POLLEN Metrology, a software editor specialized in signal processing, hybrid metrology and machine learning for the automatic processing of heterogeneous data. This funding supports a postdoc to accelerate the introduction of artificial intelligence, and in particular computer vision, techniques, into the manufacture of new generation of microprocessors. Karteek Alahari and Valerie Perrier (LJK, UGA) jointly supervise a postdoc as part of this collaboration. This collaboration ended in 2019.

9.2. National Initiatives

9.2.1. ANR Project Macaron

Participants: Julien Mairal, Zaid Harchaoui [Univ. Washington], Laurent Jacob [CNRS, LBBE Laboratory], Michael Blum [CNRS, TIMC Laboratory], Joseph Salmon [Telecom ParisTech], Mikita Dvornik, Daan Wynen.

The project MACARON is an endeavor to develop new mathematical and algorithmic tools for making machine learning more scalable. Our ultimate goal is to use data for solving scientific problems and automatically converting data into scientific knowledge by using machine learning techniques. Therefore, our project has two different axes, a methodological one, and an applied one driven by explicit problems. The methodological axis addresses the limitations of current machine learning for simultaneously dealing with large-scale data and huge models. The second axis addresses open scientific problems in bioinformatics, computer vision, image processing, and neuroscience, where a massive amount of data is currently produced, and where huge-dimensional models yield similar computational problems.

This is a 4 years and half project, funded by ANR under the program “Jeunes chercheurs, jeunes chercheuses”, which started in October 2014 and ended in March 2019. The principal investigator is Julien Mairal.

9.2.2. ANR Project DeepInFrance

Participants: Jakob Verbeek, Adria Ruiz Ovejero.

DeepInFrance (Machine learning with deep neural networks) project also aims at bringing together complementary machine learning, computer vision and machine listening research groups working on deep learning with GPUs in order to provide the community with the knowledge, the visibility and the tools that brings France among the key players in deep learning. The long-term vision of Deep in France is to open new frontiers and foster research towards algorithms capable of discovering sense in data in an automatic manner, a stepping stone before the more ambitious far-end goal of machine reasoning. The project partners are: INSA Rouen, Univ. Caen, Inria, UPMC, Aix-Marseille Univ., Univ. Nice Sophia Antipolis.

9.2.3. ANR Project AVENUE

Participant: 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.

9.3. European Initiatives

9.3.1. FP7 & H2020 Projects

9.3.1.1. ERC Advanced grant Allegro

Participants: Cordelia Schmid, Konstantin Shmelkov, Vladyslav Sydorov, Daan Wynen, Nikita Dvornik, Xavier Martin.

The ERC advanced grant ALLEGRO started in April 2013 and will end in April 2019. The aim of ALLEGRO is to automatically learn from large quantities of data with weak labels. A massive and ever growing amount of digital image and video content is available today. It often comes with additional information, such as text, audio or other meta-data, that forms a rather sparse and noisy, yet rich and diverse source of annotation, ideally suited to emerging weakly supervised and active machine learning technology. The ALLEGRO project will take visual recognition to the next level by using this largely untapped source of data to automatically learn visual models. We will develop approaches capable of autonomously exploring evolving data collections, selecting the relevant information, and determining the visual models most appropriate for different object, scene, and activity categories. An emphasis will be put on learning visual models from video, a particularly rich source of information, and on the representation of human activities, one of today’s most challenging problems in computer vision.

9.3.1.2. ERC Starting grant Solaris

Participants: Julien Mairal, Ghislain Durif, Andrei Kulunchakov, Alberto Bietti, Dexiong Chen, Gregoire Mialon.

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. International Initiatives

9.4.1. Inria International Labs

Inria@EastCoast

Associate Team involved in the International Lab:

9.4.1.1. GAYA

Title: Semantic and Geometric Models for Video Interpretation

International Partner (Institution - Laboratory - Researcher):

Carnegie Mellon University (United States) - Machine Learning Department - Katerina Fragkiadaki

Start year: 2019

See also: https://team.inria.fr/gaya/

We propose to renew the associate team GAYA, with the primary goal of interpreting videos in terms of recognizing actions, understanding the human-human and human-object interactions. In the first three years, the team has started addressing the problem of learning an efficient and robust video representation to attack this challenge. GAYA will now focus on building semantic models, wherein we learn incremental, joint audio-visual models, with limited supervision, and also geometric models, where we study the geometric properties of object shapes to better recognize them. The team consists of researchers from two Inria project-teams (Thoth and WILLOW), a US university (Carnegie Mellon University [CMU]) as the main partner team, and another US university (UC Berkeley) as a secondary partner. It will allow the partners to effectively combine their respective strengths in areas such as inference and machine learning approaches for vision tasks, joint audio-visual models, large-scale learning, geometric reasoning. The main expected outcomes of this collaboration are: new machine learning algorithms for handling minimally annotated multi-modal data, large-scale public datasets for benchmarking, theoretical analysis of objects shapes and contours. This associate team originally started in 2016, and was extended in 2019 for another 3 years.

9.4.2. Inria International Partners

9.4.2.1. Informal International Partners

- **MPI Tübingen:** Cordelia Schmid collaborates with Michael Black, a research director at MPI, starting in 2013. End of 2015 she was award a Humbolt research award funding a long-term research project with colleagues at MPI. In 2019, the project resulted in the development of an approach for object interaction [20].

9.4.3. Participation in Other International Programs

- **Indo-French project EVEREST** with IIIT Hyderabad, India, funded by CEFIPRA (Centre Franco-Indien pour la Promotion de la Recherche Avancee). The aim of this project between Cordelia Schmid, Karteek Alahari and C. V. Jawahar (IIIT Hyderabad) is to enable the use of rich, complex models that are required to address the challenges of high-level computer vision. The work plan for the project will follow three directions. First, we will develop a learning framework that can handle weak annotations. Second, we will build formulations to solve the non-convex optimization problem resulting from the learning framework. Third, we will develop efficient and accurate energy minimization algorithms, in order to make the optimization computationally feasible.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

9.5.1.1. Internships

- Pia Bideau (PhD Student, Univ. Massachusetts Amherst) was an intern in the team until Jan 2019.
- Avijit Dasgupta (PhD Student, IIIT Hyderabad, India) was an intern in the team from Feb to May 2019.
- Gunnar Sigurdsson (PhD student, CMU) was an intern in the team from Jan to Mar 2019.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Organisation

10.1.1.1. General Chair, Scientific Chair

- C. Schmid is a general chair for ECCV 2020, ICCV 2023.

10.1.1.2. Member of the Organizing Committees

- K. Alahari and J. Mairal co-organized the international summer school PAISS 2019.
- J. Mairal is a member of the organizing committee for the international conference SIAM Imaging Science 2020.
- J. Mairal is a co-organizer of the workshop OSL'19 at Les Houches.
- J. Mairal co-organized a discussion session at the Ellis/Dali workshop, San Sebastian, 2019.

10.1.2. Scientific Events: Selection

10.1.2.1. Member of the Conference Program Committees

- K. Alahari: area chair for CVPR 2020, ICCV 2019.
- K. Alahari: senior program committee member for AAAI 2020, IJCAI 2019, IJCAI 2020.
- K. Alahari: doctoral consortium chair for ICCV 2023.
- J. Mairal: area chair for NeurIPS 2019, AISTATS 2020 and ECCV 2020.
- J. Mairal: tutorial chair for CVPR 2022.
- C. Schmid: area chair for ICCV 2019.
- C. Schmid: senior area chair for NeurIPS 2019.
- J. Verbeek: area chair for ICCV 2019.

10.1.2.2. 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 AAAI, AISTATS, CVPR, ICCV, ICML, ICLR, NeurIPS in 2019.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- K. Alahari: Associate editor of the International Journal of Computer Vision, since 2019.
- K. Alahari: Associate editor for Computer Vision and Image Understanding journal, since 2018.
- 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.
- J. Mairal: Associate editor of the SIAM Journal of Imaging Science, since 2018.
- J. Verbeek: Associate editor International Journal on Computer Vision, 2014-2019.
- J. Verbeek: Associate editor IEEE Transactions Pattern Analysis and Machine Intelligence, since 2018.

10.1.3.2. 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, Machine Learning). Some of them also review for other reputed journals such as PLOS ONE, SIAM Journal on Optimization, SIAM Imaging Science.

10.1.4. Invited Talks

- K. Alahari: Speaker on the Panel on AI and Mathematics, Knowledge Summit, Lyon, France, 2019.
- K. Alahari: Invited talk, LIAMA workshop, Paris, France, 2019.
- A. Bietti: Invited talk, GIPSA Lab, Grenoble, 2019.
- A. Bietti: Invited talk, UC Berkeley, 2019.
- A. Bietti: Seminar, Microsoft Research AI, Redmond, 2019.
- A. Bietti: Seminar, TTI-Chicago, 2019.
- D. Chen: Machine Learning in Computational Biology (MLCB) workshop on recurrent kernel networks, Vancouver, 2019.
- J. Mairal: Invited talk at the YES workshop, Eindhoven, 2019.
- J. Mairal: Talk in mini-symposium, ICCOPT, Berlin, 2019.
- J. Mairal: Invited talk at the Imaging and Machine Learning conference, IHP, Paris, 2019.
- J. Mairal: Seminar. Centrale Lille, 2019.
- R. Klokov: Invited talk at Christmas Colloquium on Computer Vision, Yandex, Moscow, 2019.
- C. Schmid: Invited speaker at BMVA symposium in Video Understanding, London, September 2019.
- C. Schmid: Keynote speaker at BMVC, Cardiff, UK, September 2019.
- C. Schmid: Keynote speaker at SIGIR, Paris, July 2019.
- C. Schmid: Invited speaker at Computer Vision after 5 Years, in conjunction with CVPR, June 2019.
- C. Schmid: Invited speaker at Tutorial on Unifying Human Activity Understanding, in conjunction with CVPR, June 2019.
- C. Schmid: Invited speaker at Facebook AI Video Summit, June 2019.
- C. Schmid: Keynote speaker at AI Experts Workshop in conjunction with the AI for Good Global Summit, Geneva, May 2019.
- C. Schmid: Invited speaker at Women in Data Science Conference, Zürich, April 2019.
- C. Schmid: Invited speaker at Collège de France seminar (chair of Stephane Mallat), February 2019.
- C. Schmid: Talk at Google EMEA research days, Zurich, December 2019.

- C. Schmid: Talk at Workshop on AI for Robotics, Naver, Grenoble, November 2019.
- C. Schmid: Talk at Workshop, Robotics: A Challenge for the Artificial Intelligence, Toulouse, October 2019.
- C. Schmid: Presentation at PRAIRIE inauguration, Paris, October 2019.
- C. Schmid: Seminar at DeepMind, London, September 2019.
- C. Schmid: Seminar at Intel Network on Intelligent Systems, Munich, September 2019.
- C. Schmid: Seminar at Ellis workshop, September 2019.
- C. Schmid: Seminar at MPI Tübingen, July 2019.
- C. Schmid: Seminar at WILLOW/SIERRA retreat, Marseille, June 2019.
- C. Schmid: Dinner speaker at the workshop “Women in Computer Vision”, in conjunction with CVPR’19.
- C. Schmid: Seminar at Google MTV, April 2019.
- C. Schmid: Seminar at ETH Zürich, March 2019.
- J. Verbeek: Invited talk at Breaking the Surface Workshop on maritime robotics and its applications, Biograd na Moru, Croatia, Oct 2019.
- J. Verbeek: Invited talk at Dagstuhl Workshop on Joint Processing of Language and Visual Data for Better Automated Understanding, Germany, Jan 2019.
- D. Wynen: SMILE Reading Group Paris, 2019.

10.1.5. Leadership within the Scientific Community

- J. Mairal, J. Verbeek and C. Schmid became Ellis fellows.
- C. Schmid: Participation in a round table on AI, a technology for innovation, forum 5i, Grenoble, May 2019.
- C. Schmid: Animating several mentorship sessions at Women in Data Science Conference, Zürich, April 2019.
- C. Schmid: Mentor at the Doctoral Consortium, in conjunction with ICCV’19, CVPR’19.
- C. Schmid: Mentor for female PhD students at the workshop “Women in Computer Vision", CVPR’19.

10.1.6. Scientific Expertise

- J. Mairal: Judge for the IBM Watson AI Xprize.
- J. Mairal: Expert for ANR.

10.1.7. Research Administration

- K. Alahari: One of the two referents for Human Resources - Excellence in Research (HRS4R) at Inria Grenoble.
- J. Mairal: Jury member for the Inria starting and advanced research positions.
- C. Schmid: Member of Scientific Advisory Committee of the Helmholtz AI Cooperation Unit, 2020—
- C. Schmid: Member of scientific advisory board for the German Competence Centers for AI Research, 2019—
- J. Verbeek: Member steering committee MinaLogic, innovation cluster for digital technologies based in France’s Auvergne-Rhône-Alpes region, 2018-2019.
- J. Verbeek: Scientific correspondent national project calls, Inria Grenoble, 2017-2019.
- J. Verbeek: Member Scientific council Advanced Data Mining axis of Persyval Laboratory of Excellence, Grenoble, 2015-2019.

- J. Verbeek: Member Inria Grenoble working group on HPC - Big Data - Machine learning, 2018-2019.
- J. Verbeek: Member of Inria Commission Administrative Paritaire (advises on matters about individual careers: such as promotions, temporary outsourcing, etc.), 2016-2019.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

- Doctorat: K. Alahari, Lecturer at the CVIT summer school on machine learning, 4h eqTD, IIIT Hyderabad, India.
- Doctorat: J. Mairal, Large-Scale Optimization for Machine Learning, 9h eqTD, Lecture at OBA summer school, Veroli, 2019.
- Doctorat: J. Mairal, Large-Scale Optimization for Machine Learning, 4.5h eqTD. Invited Tutorial at IEEE Data Science Workshop on Minneapolis, 2019.
- Doctorat: J. Mairal, Large-Scale Optimization for Machine Learning, 4.5h eqTD. Invited Tutorial at the Quantitative BioImaging Conference, Rennes, 2019.
- Doctorat: C. Schmid, Course on action recognition at Data Science Summer School, Paris, June 2019.
- Doctorat: C. Schmid, Course on action recognition at Prairie artificial intelligence summer school (PAISS), 2.25h eqTD, Paris, October 2019.
- Master: K. Alahari, C. Schmid, Object recognition, Master-2 Computer Science, Grenoble University, 15.75h eqTD together, 2019.
- Master: K. Alahari, Understanding Big Visual Data, 13.5h eqTD, M2, Grenoble INP, France.
- Master: K. Alahari, Graphical Models Inference and Learning, 18h eqTD, M2, CentraleSupelec, Paris, France.
- Master: K. Alahari, Introduction to computer vision, 9h eqTD, M1, ENS Paris, France.
- Master: J. Mairal, Kernel methods for statistical learning, 15h eqTD, M2, Ecole Normale Supérieure, Cachan, France.
- Master: J. Mairal, Advanced Learning Models, 13.5h eqTD, M2, UGA, Grenoble.
- Master: C. Schmid, Object recognition and computer vision, Master-2 MVA, ENS, 9h eqTD, 2019.
- Master: A. Sablayrolles, Fundamentals of Machine Learning, African Masters of Machine Intelligence, Kigali, Rwanda.

10.2.2. Supervision

HDR: Karteek Alahari, Human, Motion and Other Priors for Partially-Supervised Recognition, Univ. Grenoble Alpes, 28/1/2019.

PhD: Alberto Bietti, Foundations of deep convolutional models through kernel methods, Univ. Grenoble Alpes, 27/11/2019, director: Julien Mairal.

PhD: Nikita Dvornik, Learning with Limited Annotated Data for Visual Understanding, Univ. Grenoble Alpes, 26/11/2019, thesis directors: Cordelia Schmid and Julien Mairal.

PhD: Konstantin Shmelkov, Approaches for incremental learning and image generation, Univ. Grenoble Alpes, 29/3/2019, thesis directors: Karteek Alahari and Cordelia Schmid.

10.2.3. Juries

- K. Alahari: External examiner for the PhD thesis of Alessandro di Martino, University of Bath, UK.
- K. Alahari: Examiner for the PhD thesis of Thomas Robert, Sorbonne Université, Paris, France.
- K. Alahari: Examiner for the PhD thesis of D. Khuê Lê-Huu, Université Paris-Saclay, France.

- K. Alahari: Member of comité de suivi for the PhD thesis of Miguel Angel Solinas, Univ. Grenoble-Alpes, France.
- J. Mairal: Reviewer for the PhD thesis of Zhenyu Liao, Université Paris-Saclay.
- J. Mairal: Reviewer for the PhD thesis of Belhal Karimi, Université Paris-Saclay
- J. Mairal: Reviewer for the PhD thesis of Martin Bompaire, Université Paris-Saclay
- J. Mairal: Reviewer for the PhD thesis of Yassine Yaakoubi, Polytechnique Montréal.
- J. Mairal: Examinateur for the PhD thesis of Mathurin Massias, Université Paris-Saclay.
- J. Mairal: Member of comité de suivi for the PhD thesis of Olga Permiakova, Univ. Grenoble Alpes.
- J. Verbeek: Member supervisory commitee for PhD of Riccardo Del Chiaro, 2018-2020, Univ. Florence, Italy.
- J. Verbeek: Member supervisory commitee for PhD of Fabien Baradel, 2017-2019, INSA Lyon, France.
- J. Verbeek: External reviewer for Shell Xu Hu, 2019, Ecole des Ponts, Paris Tech, Univ. Paris Est, Paris, France.
- J. Verbeek: Rapporteur for Hedi Ben Younes, 2019, Sorbonne University, Paris, France.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] K. Alahari. *Human, Motion and Other Priors for Partially-Supervised Recognition*, Communauté Université Grenoble Alpes, January 2019, Habilitation à diriger des recherches, https://hal.inria.fr/tel-02269024

[2] K. Shmelkov. *Approaches for incremental learning and image generation*, Université Grenoble Alpes, March 2019, https://tel.archives-ouvertes.fr/tel-02183259

Articles in International Peer-Reviewed Journals

[3] A. Bietti, J. Mairal. *Group Invariance, Stability to Deformations, and Complexity of Deep Convolutional Representations*, in "Journal of Machine Learning Research", 2019, vol. 20, n^{o} 1, pp. 1-49, https://arxiv.org/abs/1706.03078 , https://hal.inria.fr/hal-01536004

[4] D. Chen, L. Jacob, J. Mairal. *Biological Sequence Modeling with Convolutional Kernel Networks*, in "Bioinformatics", September 2019, vol. 35, n^{o} 18, pp. 3294–3302 [*DOI :* 10.1093/bioinformatics/btz094], https://hal.inria.fr/hal-01632912

[5] D. Derkach, A. Ruiz, F. M. Sukno. *Tensor Decomposition and Non-linear Manifold Modeling for 3D Head Pose Estimation*, in "International Journal of Computer Vision", October 2019, vol. 127, n^{o} 10, pp. 1565-1585 [*DOI :* 10.1007/s11263-019-01208-x], https://hal.archives-ouvertes.fr/hal-02267568

[6] G. Durif, L. Modolo, J. E. Mold, S. Lambert-Lacroix, F. Picard. *Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis*, in "Bioinformatics", October 2019, vol. 20, pp. 4011–4019, https://arxiv.org/abs/1710.11028 [*DOI :* 10.1093/bioinformatics/btz177], https://hal.archives-ouvertes.fr/hal-01649275

[7] N. DVORNIK, J. MAIRAL, C. SCHMID. *On the Importance of Visual Context for Data Augmentation in Scene Understanding*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", December 2019, pp. 1-15, forthcoming [*DOI :* 10.1109/TPAMI.2019.2961896], https://hal.archives-ouvertes.fr/hal-01869784

[8] H. LIN, J. MAIRAL, Z. HARCHAOUI. *An Inexact Variable Metric Proximal Point Algorithm for Generic Quasi-Newton Acceleration*, in "SIAM Journal on Optimization", May 2019, vol. 29, n^{o} 2, pp. 1408-1443, https://arxiv.org/abs/1610.00960 [*DOI :* 10.1137/17M1125157], https://hal.inria.fr/hal-01376079

[9] G. ROGEZ, P. WEINZAEPFEL, C. SCHMID. *LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2019, pp. 1-15, forthcoming [*DOI :* 10.1109/TPAMI.2019.2892985], https://hal.archives-ouvertes.fr/hal-01961189

[10] P. TOKMAKOV, C. SCHMID, K. ALAHARI. *Learning to Segment Moving Objects*, in "International Journal of Computer Vision", March 2019, vol. 127, n^{o} 3, pp. 282–301, https://arxiv.org/abs/1712.01127 [*DOI :* 10.1007/s11263-018-1122-2], https://hal.archives-ouvertes.fr/hal-01653720

International Conferences with Proceedings

[11] A. BIETTI, J. MAIRAL. *On the Inductive Bias of Neural Tangent Kernels*, in "NeurIPS 2019 - Thirty-third Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, pp. 1-24, https://arxiv.org/abs/1905.12173 , https://hal.inria.fr/hal-02144221

[12] A. BIETTI, G. MIALON, D. CHEN, J. MAIRAL. *A Kernel Perspective for Regularizing Deep Neural Networks*, in "ICML 2019 - 36th International Conference on Machine Learning", Long Beach, United States, Proceedings of Machine Learning Research, June 2019, vol. 97, pp. 664-674, https://arxiv.org/abs/1810.00363 , https://hal.inria.fr/hal-01884632

[13] M. CARON, P. BOJANOWSKI, J. MAIRAL, A. JOULIN. *Unsupervised Pre-Training of Image Features on Non-Curated Data*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, Proceedings of the International Conference on Computer Vision (ICCV), October 2019, pp. 1-10, https://hal.archives-ouvertes.fr/hal-02119564

[14] D. CHEN, L. JACOB, J. MAIRAL. *Biological Sequence Modeling with Convolutional Kernel Networks*, in "RECOMB 2019 - 23rd Annual International Conference Research in Computational Molecular Biology", Washington DC, United States, Springer, May 2019, pp. 1-2 [*DOI :* 10.1007/978-3-030-17083-7], https://hal.archives-ouvertes.fr/hal-02388776

[15] D. CHEN, L. JACOB, J. MAIRAL. *Recurrent Kernel Networks*, in "NeurIPS 2019 - Thirty-third Conference Neural Information Processing Systems", Vancouver, Canada, December 2019, pp. 1-19, https://arxiv.org/abs/1906.03200 , https://hal.inria.fr/hal-02151135

[16] N. CRASTO, P. WEINZAEPFEL, K. ALAHARI, C. SCHMID. *MARS: Motion-Augmented RGB Stream for Action Recognition*, in "CVPR 2019 - IEEE Conference on Computer Vision & Pattern Recognition", Long Beach, CA, United States, IEEE, June 2019, pp. 1-10, https://hal.inria.fr/hal-02140558

[17] N. DVORNIK, C. SCHMID, J. MAIRAL. *Diversity with Cooperation: Ensemble Methods for Few-Shot Classification*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, pp. 1-12, https://arxiv.org/abs/1903.11341 - Added experiments with different network architectures and input image resolutions, https://hal.archives-ouvertes.fr/hal-02080004

[18] M. Elbayad, J. Gu, E. Grave, M. Auli. *Depth-adaptive Transformer*, in "ICLR 2020 - Eighth International Conference on Learning Representations", Addis Ababa, Ethiopia, December 2019, pp. 1-14, https://hal.inria.fr/hal-02422914

[19] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, G. Rogez. *Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, pp. 1-10, https://hal.inria.fr/hal-02242795

[20] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, C. Schmid. *Learning joint reconstruction of hands and manipulated objects*, in "CVPR 2019 - IEEE Conference on Computer Vision and Pattern Recognition", Long Beach, United States, IEEE, June 2019, pp. 1-14, https://hal.archives-ouvertes.fr/hal-02429093

[21] R. Klokov, J. Verbeek, E. Boyer. *Probabilistic Reconstruction Networks for 3D Shape Inference from a Single Image*, in "BMVC 2019 - British Machine Vision Conference", Cardiff, United Kingdom, September 2019, pp. 1-15, https://arxiv.org/abs/1908.07475 - Awarded with Best Science Paper Honourable Mention Award at BMVC'19., https://hal.inria.fr/hal-02268466

[22] A. Kulunchakov, J. Mairal. *A Generic Acceleration Framework for Stochastic Composite Optimization*, in "NeurIPS 2019 - Thirty-third Conference Neural Information Processing Systems", Vancouver, Canada, December 2019, pp. 1-24, https://arxiv.org/abs/1906.01164 , https://hal.inria.fr/hal-02139489

[23] A. Kulunchakov, J. Mairal. *Estimate Sequences for Variance-Reduced Stochastic Composite Optimization*, in "ICML 2019 - 36th International Conference on Machine Learning", Long Beach, United States, June 2019, pp. 1-24, https://arxiv.org/abs/1905.02374 - short version of preprint arXiv:1901.08788, https://hal.inria.fr/hal-02121913

[24] T. Lucas, K. Shmelkov, K. Alahari, C. Schmid, J. Verbeek. *Adaptive Density Estimation for Generative Models*, in "NeurIPS 2019 - Thirty-third Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, pp. 1-24, https://hal.archives-ouvertes.fr/hal-01886285

[25] A. Pashevich, R. Strudel, I. Kalevatykh, I. Laptev, C. Schmid. *Learning to Augment Synthetic Images for Sim2Real Policy Transfer*, in "IROS 2019 - IEEE/RSJ International Conference on Intelligent Robots and Systems", Macao, China, November 2019, pp. 1-6, https://arxiv.org/abs/1903.07740 - 7 pages, https://hal.archives-ouvertes.fr/hal-02273326

[26] J. Peyre, I. Laptev, C. Schmid, J. Sivic. *Detecting unseen visual relations using analogies*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, https://arxiv.org/abs/1812.05736v3 , https://hal.archives-ouvertes.fr/hal-01975760

[27] A. Ruiz, J. Verbeek. *Adaptative Inference Cost With Convolutional Neural Mixture Models*, in "ICCV 2019 - International Conference on Computer Vision", Seoul, South Korea, October 2019, pp. 1-12, https://hal.archives-ouvertes.fr/hal-02267564

[28] A. Sablayrolles, M. Douze, Y. Ollivier, C. Schmid, H. Jégou. *White-box vs Black-box: Bayes Optimal Strategies for Membership Inference*, in "ICML 2019 - 36th International Conference on Machine Learning", Long Beach, United States, June 2019, https://arxiv.org/abs/1908.11229 , https://hal.inria.fr/hal-02278902

[29] A. SABLAYROLLES, M. DOUZE, C. SCHMID, H. JÉGOU. *Spreading vectors for similarity search*, in "ICLR 2019 - 7th International Conference on Learning Representations", New Orleans, United States, May 2019, pp. 1-13, https://arxiv.org/abs/1806.03198 - Published at ICLR 2019, https://hal.inria.fr/hal-02278905

[30] V. SYDOROV, K. ALAHARI, C. SCHMID. *Focused Attention for Action Recognition*, in "BMVC 2019 - British Machine Vision Conference", Cardiff, United Kingdom, September 2019, pp. 1-12, https://hal.archives-ouvertes.fr/hal-02292339

[31] M. VLADIMIROVA, J. VERBEEK, P. MESEJO, J. ARBEL. *Understanding Priors in Bayesian Neural Networks at the Unit Level*, in "ICML 2019 - 36th International Conference on Machine Learning", Long Beach, United States, Proceedings of the 36th International Conference on Machine Learning, June 2019, vol. 97, pp. 6458-6467, https://arxiv.org/abs/1810.05193 - 10 pages, 5 figures, ICML'19 conference [*DOI :* 10.05193], https://hal.archives-ouvertes.fr/hal-02177151

Conferences without Proceedings

[32] A. RUIZ, O. MARTINEZ, X. BINEFA, J. VERBEEK. *Learning Disentangled Representations with Reference-Based Variational Autoencoders*, in "ICLR workshop on Learning from Limited Labeled Data", New Orleans, United States, May 2019, pp. 1-17, https://hal.inria.fr/hal-01896007

Other Publications

[33] G. CHÉRON, A. OSOKIN, I. LAPTEV, C. SCHMID. *Modeling Spatio-Temporal Human Track Structure for Action Localization*, January 2019, https://arxiv.org/abs/1806.11008 - working paper or preprint, https://hal.inria.fr/hal-01979583

[34] A. ISCEN, G. TOLIAS, Y. AVRITHIS, O. CHUM, C. SCHMID. *Graph Convolutional Networks for Learning with Few Clean and many Noisy Labels*, November 2019, https://arxiv.org/abs/1910.00324 - working paper or preprint [*DOI :* 10.00324], https://hal.inria.fr/hal-02370212

[35] A. KULUNCHAKOV, J. MAIRAL. *Estimate Sequences for Stochastic Composite Optimization: Variance Reduction, Acceleration, and Robustness to Noise*, January 2019, https://arxiv.org/abs/1901.08788 - working paper or preprint, https://hal.inria.fr/hal-01993531

[36] B. LECOUAT, J. PONCE, J. MAIRAL. *Revisiting Non Local Sparse Models for Image Restoration*, December 2019, working paper or preprint, https://hal.inria.fr/hal-02414291

[37] X. LI, S. WANG, Y. ZHAO, J. VERBEEK, J. KANNALA. *Hierarchical Scene Coordinate Classification and Regression for Visual Localization*, November 2019, https://arxiv.org/abs/1909.06216 - working paper or preprint, https://hal.inria.fr/hal-02384675

[38] J. MAIRAL. *Cyanure: An Open-Source Toolbox for Empirical Risk Minimization for Python, C++, and soon more*, December 2019, working paper or preprint, https://hal.inria.fr/hal-02417766

[39] G. MIALON, A. D'ASPREMONT, J. MAIRAL. *Screening Data Points in Empirical Risk Minimization via Ellipsoidal Regions and Safe Loss Functions*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02395624

[40] R. STRUDEL, A. PASHEVICH, I. KALEVATYKH, I. LAPTEV, J. SIVIC, C. SCHMID. *Learning to combine primitive skills: A step towards versatile robotic manipulation*, August 2019, https://arxiv.org/abs/1908.00722 - 11 pages, https://hal.archives-ouvertes.fr/hal-02274969

[41] G. VAROL, I. LAPTEV, C. SCHMID, A. ZISSERMAN. *Synthetic Humans for Action Recognition from Unseen Viewpoints*, January 2020, https://arxiv.org/abs/1912.04070 - working paper or preprint, https://hal.inria.fr/hal-02435731