Activity Report 2018

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

Post-Doctoral Fellows

Henrique Morimitsu [Inria, until Feb 2018]
Adria Ruiz Ovejero [Inria, from Jan 2018]

PhD Students

Minttu Alakuijala [Inria, from Nov 2018]
Alberto Bietti [Inria]
Mathilde Caron [Inria, from Apr until Sep 2018, Facebook CIFRE, from Oct 2018]
Dexiong Chen [Univ. Paris-Saclay]
Guilhem Cheron [Inria, until Dec 2018]
Mikita Dvornik [Inria]
Maha Elbayad [Univ. Grenoble Alpes]
Valentin Gabeur [Inria, from Oct 2018]
Yana Hasson [Inria]
Ekaterina Iakovleva [Inria, from Oct 2018]
Roman Klokov [Inria]
Andrei Kulunchakov [Inria]

Pauline Luc [Facebook CIFRE]
Thomas Lucas [Univ. Grenoble Alpes]
Grégoire Mialon [Inria, from Oct 2018]
Alexander Pashevich [Inria]
Alexandre Sablayrolles [Facebook CIFRE]
Konstantin Shmelkov [Inria]
Vladyslav Sydorov [Inria]
Pavel Tokmakov [Inria, until Jan 2018]
Gul Varol [Inria]
Nitika Verma [Univ. Grenoble Alpes]
Daan Wynen [Inria]

Technical staff

Ghislain Durif [Inria]
François Gindraud [Inria, from Oct 2018]
Xavier Martin [Inria]

Interns

Michel Aractingi [Inria, from Feb 2018 until Jul 2018]
Vasileios Choutas [Inria, until Mar 2018]
Nieves Crasto [Inria, from Feb 2018 until Nov 2018]
Valentin Gabeur [Inria, from Apr 2018 until Sep 2018]
Mariia Garkavenko [Inria, from Feb 2018 until Jul 2018]
Ekaterina Iakovleva [Inria, from Feb 2018 until Jul 2018]
Robin Maillot [CEA, from Mar 2018 until Jul 2018]
Grégoire Mialon [Inria, from Apr 2018 until Sep 2018]

Administrative Assistant

Nathalie Gillot [Inria]

Visiting Scientist

Pia Bideau [Univ. Massachusetts Amherst, from Sep 2018 until Dec 2018]

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 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

[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.

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 leader-boards, 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

- Alberto Bietti received the Jean-Claude Dodu 2018 prize at Journees SMAI-MODE, Autrans.
- Pauline Luc was one of the top-200 reviewers at NeurIPS 2018.
- Grégory Rogez and Cordelia Schmid received an Amazon Academic Research Award.
- Cordelia Schmid received the Koenderink prize for fundamental contributions in computer vision that have withstood the test of time at ECCV 2018.

5.1.2. Dissemination

- The team co-organized PAISS 2018, an international AI summer school in Grenoble. This event brought together 200 participants representing 44 different nationalities. The participants were selected from 700 applications, with 60% students, 15% academics, and 25% industrials. 25% of these participants were women.

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
- Contact: Grégory Rogez
- 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: CKN-Seq is a software package for predicting transcription factor binding sites. It was shipped with the BiorXiv preprint

D. Chen, L. Jacob, and J. Mairal. Predicting Transcription Factor Binding Sites with Convolutional Kernel Networks. 2017.

The software is implemented in PyTorch.

- 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. Loter

Loter: A software package to infer local ancestry for a wide range of species
KEYWORDS: Local Ancestry Inference - Bioinformatics
SCIENTIFIC DESCRIPTION: Admixture between populations provides opportunity to study biological adaptation and phenotypic variation. Admixture studies can rely on local ancestry inference for admixed individuals, which consists of computing at each locus the number of copies that originate from ancestral source populations. Loter is a software package that does not require any biological parameter besides haplotype data in order to make local ancestry inference available for a wide range of species.
FUNCTIONAL DESCRIPTION: Loter is a Python package for haplotype phasing and local ancestry inference.

NEWS OF THE YEAR: The software package was shipped with the biorxiv preprint T. Dias-Alves, J. Mairal, and M. Blum. Loter: A Software Package to Infer Local Ancestry for a Wide Range of Species. preprint BiorXiv. 2017

- Participants: Thomas Dias-Alves, Michael Blum and Julien Mairal
- Partners: UGA - CNRS
- Contact: Julien Mairal
- Publication: Loter: A software package to infer local ancestry for a wide range of species
- URL: https://github.com/bcm-uga/Loter

6.4. SPAMS

SPArse Modeling Software
KEYWORDS: Signal processing - Machine learning
FUNCTIONAL DESCRIPTION: SPAMS is an open-source software package for sparse estimation
NEWS OF THE YEAR: The version 2.6.1 of the software package is now compatible with Python v3, R v3, comes with pre-compiled Matlab packages, and is now available on the conda and PyPi package managers.

- Participants: Ghislain Durif and Julien Mairal
- Contact: Julien Mairal
- URL: http://spams-devel.gforge.inria.fr/

6.5. 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

- Participants: Karteek Alahari, Cordelia Schmid and Pavel Tokmakov
- Contact: Pavel Tokmakov
- Publication: Learning Video Object Segmentation with Visual Memory
- URL: http://lear.inrialpes.fr/research/lvo/

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

6.7. 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

7. New Results

7.1. Visual Recognition in Images and Videos

7.1.1. Actor and Observer: Joint Modeling of First and Third-Person Videos

Participants: Gunnar Sigurdsson [CMU], Abhinav Gupta [CMU], Cordelia Schmid, Ali Farhadi [AI2, Univ. Washington], Karteek Alahari.

Several theories in cognitive neuroscience suggest that when people interact with the world, or simulate interactions, they do so from a first-person egocentric perspective, and seamlessly transfer knowledge between third-person (observer) and first-person (actor). Despite this, learning such models for human action recognition has not been achievable due to the lack of data. Our work in [36] takes a step in this direction, with the introduction of Charades-Ego, a large-scale dataset of paired first-person and third-person videos, involving 112 people, with 4000 paired videos. This enables learning the link between the two, actor and observer perspectives. Thereby, we address one of the biggest bottlenecks facing egocentric vision research, providing a link from first-person to the abundant third-person data on the web. We use this data to learn a joint representation of first and third-person videos, with only weak supervision, and show its effectiveness for transferring knowledge from the third-person to the first-person domain.

7.1.2. Learning to Segment Moving Objects

Participants: Pavel Tokmakov, Cordelia Schmid, Karteek Alahari.

We study the problem of segmenting moving objects in unconstrained videos [17]. Given a video, the task is to segment all the objects that exhibit independent motion in at least one frame. We formulate this as a learning problem and design our framework with three cues: (i) independent object motion between a pair of frames, which complements object recognition, (ii) object appearance, which helps to correct errors in motion estimation, and (iii) temporal consistency, which imposes additional constraints on the segmentation. The framework is a two-stream neural network with an explicit memory module. The two streams encode appearance and motion cues in a video sequence respectively , while the memory module captures the evolution of objects over time, exploiting the temporal consistency. The motion stream is a convolutional neural network trained on synthetic videos to segment independently moving objects in the optical flow field. 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. For every pixel in a frame of a test video, our approach assigns an object or background label based on the learned spatio-temporal features as well as the 'visual memory' specific to the video. We evaluate our method extensively on three benchmarks, DAVIS, Freiburg-Berkeley motion seg-mentation dataset and SegTrack. In addition, we provide an extensive ablation study to investigate both the choice of the training data and the influence of each component in the proposed framework. An overview of our model is shown in Figure 1.

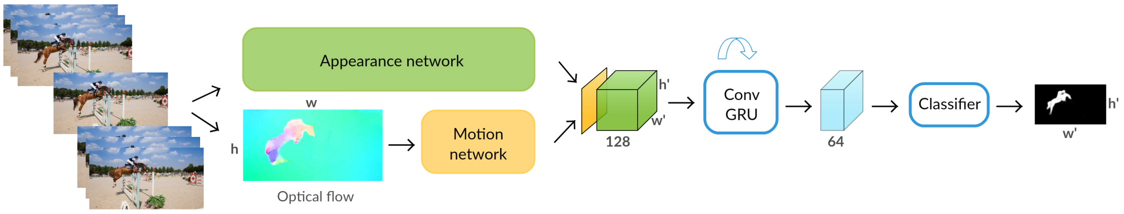


Figure 1. Overview of our segmentation approach [17]. Each video frame is processed by the appearance (green) and the motion (yellow) networks to produce an intermediate two-stream representation. The ConvGRU module combines this with the learned visual memory to compute the final segmentation result. The width (w') and height (h') of the feature map and the output are $w/8$ and $h/8$ respectively.

7.1.3. Unsupervised Learning of Artistic Styles with Archetypal Style Analysis

Participants: Daan Wynen, Cordelia Schmid, Julien Mairal.

In [39], we introduce an unsupervised learning approach to automatically discover, summarize, and manipulate artistic styles from large collections of paintings. Our method (summarized in Figure 2) is based on archetypal analysis, which is an unsupervised learning technique akin to sparse coding with a geometric interpretation. When applied to neural style representations from a collection of artworks, it learns a dictionary of archetypal styles, which can be easily visualized. After training the model, the style of a new image, which is characterized by local statistics of deep visual features, is approximated by a sparse convex combination of archetypes. This enables us to interpret which archetypal styles are present in the input image, and in which proportion. Finally, our approach allows us to manipulate the coefficients of the latent archetypal decomposition, and achieve various special effects such as style enhancement, transfer, and interpolation between multiple archetypes.

7.1.4. Learning from Web Videos for Event Classification

Participants: Nicolas Chesneau, Karteek Alahari, Cordelia Schmid.

Traditional approaches for classifying event videos rely on a manually curated training dataset. While this paradigm has achieved excellent results on benchmarks such as TrecVid multimedia event detection (MED) challenge datasets, it is restricted by the effort involved in careful annotation. Recent approaches have

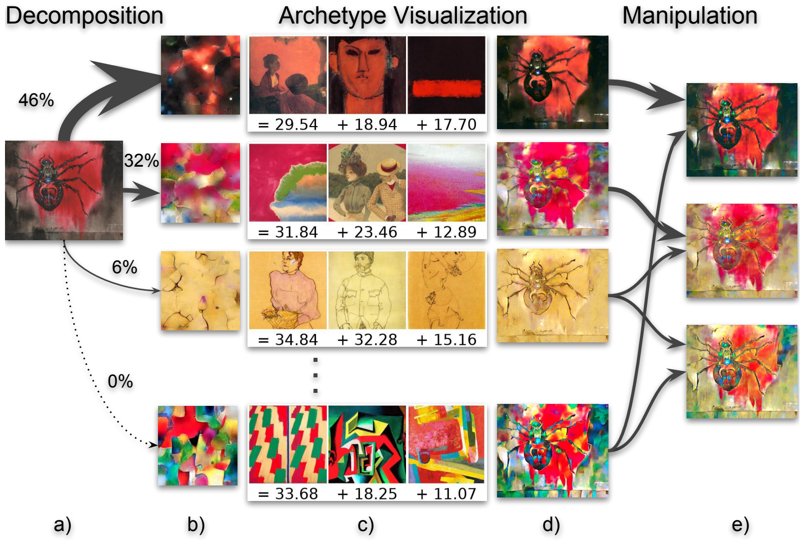


Figure 2. Using deep archetypal style analysis, we can represent the style of an artwork (a) as a convex combination of archetypes. The archetypes can be visualized as synthesized textures (b), as a convex combination of artworks (c) or, when analyzing a specific artwork, as stylized versions of that artwork itself (d). Free recombination of the archetypal styles then allows for novel stylizations of the input.

attempted to address the need for annotation by automatically extracting images from the web, or generating queries to retrieve videos. In the former case, they fail to exploit additional cues provided by video data, while in the latter, they still require some manual annotation to generate relevant queries. We take an alternate approach in [6], leveraging the synergy between visual video data and the associated textual metadata, to learn event classifiers without manually annotating any videos. Specifically, we first collect a video dataset with queries constructed automatically from textual description of events, prune irrelevant videos with text and video data, and then learn the corresponding event classifiers. We evaluate this approach in the challenging setting where no manually annotated training set is available, i.e., EK0 in the TrecVid challenge, and show state-of-the-art results on MED 2011 and 2013 datasets.

7.1.5. How good is my GAN?

Participants: Konstantin Shmelkov, Cordelia Schmid, Karteek Alahari.

Generative adversarial networks (GANs) are one of the most popular methods for generating images today. While impressive results have been validated by visual inspection, a number of quantitative criteria have emerged only recently. We argue here that the existing ones are insufficient and need to be in adequation with the task at hand. In [35] introduce two measures based on image classification—GAN-train and GAN-test (illustrated in Figure 3), which approximate the recall (diversity) and precision (quality of the image) of GANs respectively. We evaluate a number of recent GAN approaches based on these two measures and demonstrate a clear difference in performance. Furthermore, we observe that the increasing difficulty of the dataset, from CIFAR10 over CIFAR100 to ImageNet, shows an inverse correlation with the quality of the GANs, as clearly evident from our measures.

7.1.6. Modeling Visual Context is Key to Augmenting Object Detection Datasets

Participants: Nikita Dvornik, Julien Mairal, Cordelia Schmid.

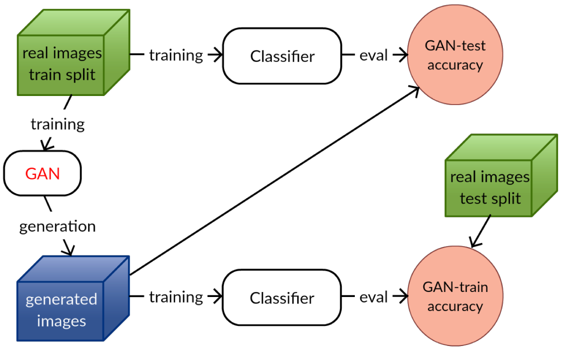


Figure 3. Illustration of GAN-train and GAN-test. GAN-train learns a classifier on GAN generated images and measures the performance on real test images. This evaluates the diversity and realism of GAN images. GAN-test learns a classifier on real images and evaluates it on GAN images. This measures how realistic GAN images are.

Performing data augmentation for learning deep neural networks is well known to be important for training visual recognition systems. By artificially increasing the number of training examples, it helps reducing overfitting and improves generalization. For object detection, classical approaches for data augmentation consist of generating images obtained by basic geometrical transformations and color changes of original training images. In [26], we go one step further and leverage segmentation annotations to increase the number of object instances present on training data. For this approach to be successful, we show that modeling appropriately the visual context surrounding objects is crucial to place them in the right environment. Otherwise, we show that the previous strategy actually hurts. Clear difference between the two approaches can is presented in Figure 4. With our context model, we achieve significant mean average precision improvements when few labeled examples are available on the VOC'12 benchmark.

7.1.7. On the Importance of Visual Context for Data Augmentation in Scene Understanding

Participants: Nikita Dvornik, Julien Mairal, Cordelia Schmid.

Performing data augmentation for learning deep neural networks is known to be important for training visual recognition systems. By artificially increasing the number of training examples, it helps reducing overfitting and improves generalization. While simple image transformations such as changing color intensity or adding random noise can already improve predictive performance in most vision tasks, larger gains can be obtained by leveraging task-specific prior knowledge. In [43], we consider object detection and semantic segmentation and augment the training images by blending objects in existing scenes, using instance segmentation annotations. We observe that randomly pasting objects on images hurts the performance, unless the object is placed in the right context. To resolve this issue, we propose an explicit context model by using a convolutional neural network, which predicts whether an image region is suitable for placing a given object or not. In our experiments, we show that by using copy-paste data augmentation with context guidance we are able to improve detection and segmentation on the PASCAL VOC12 and COCO datasets, with significant gains when few labeled examples are available. The way to augment for different tasks and annotations is presented in Figure 5. We also show that the method is not limited to datasets that come with expensive pixel-wise instance annotations and can be used when only bounding box annotations are available, by employing weakly-supervised learning for instance masks approximation.

7.1.8. Predicting future instance segmentation by forecasting convolutional features

Participants: Pauline Luc, Camille Couprie [Facebook AI Research], Yann Lecun [Facebook AI Research], Jakob Verbeek.

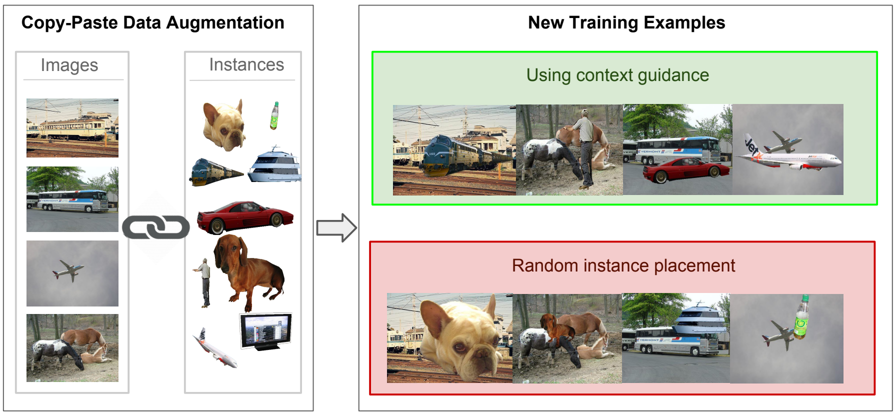


Figure 4. ***Examples of data-augmented training examples produced by our approach.*** *Images and objects are taken from the VOC'12 dataset that contains segmentation annotations. We compare the output obtained by pasting the objects with our context model vs. those obtained with random placements. Even though the results are not perfectly photorealistic and display blending artefacts, the visual context surrounding objects is more often correct with the explicit context model.*

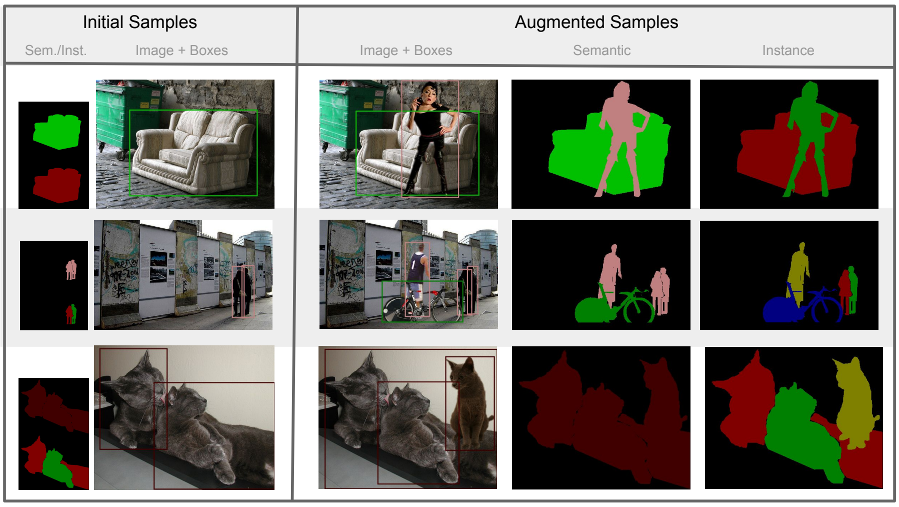


Figure 5. ***Data augmentation for different types of annotations.*** *The first column contains samples from the training dataset with corresponding semantic/instance segmentation and bounding box annotations. Columns 2-4 present the result of applying context-driven augmentation to the initial sample with corresponding annotations.*

Anticipating future events is an important prerequisite towards intelligent behavior. Video forecasting has been studied as a proxy task towards this goal. Recent work has shown that to predict semantic segmentation of future frames, forecasting at the semantic level is more effective than forecasting RGB frames and then segmenting these. In [31], we consider the more challenging problem of future instance segmentation, which additionally segments out individual objects. To deal with a varying number of output labels per image, we develop a predictive model in the space of fixed-sized convolutional features of the Mask R-CNN instance segmentation model. We apply the "detection head" of Mask R-CNN on the predicted features to produce the instance segmentation of future frames. Experiments show that this approach significantly improves over strong baselines based on optical flow and repurposed instance segmentation architectures. We show an overview of the proposed method in Figure 6.

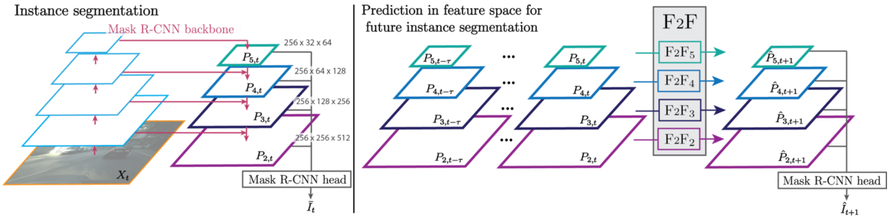


Figure 6. For future instance segmentation, we extract a pyramid of features from frames $t - \tau$ to t, and use them to predict the pyramid features for frame $t + 1$. We learn separate feature-to-feature prediction models for each level of the pyramid. The predicted features are then given as input to a downstream network to produce future instance segmentation.

7.1.9. Joint Future Semantic and Instance Segmentation Prediction

Participants: Camille Couprie [Facebook AI Research], Pauline Luc, Jakob Verbeek.

The ability to predict what will happen next from observing the past is a key component of intelligence. Methods that forecast future frames were recently introduced towards better machine intelligence. However, predicting directly in the image color space seems an overly complex task, and predicting higher level representations using semantic or instance segmentation approaches were shown to be more accurate. In [23], we introduce a novel prediction approach that encodes instance and semantic segmentation information in a single representation based on distance maps. Our graph-based modeling of the instance segmentation prediction problem allows us to obtain temporal tracks of the objects as an optimal solution to a watershed algorithm. Our experimental results on the Cityscapes dataset present state-of-the-art semantic segmentation predictions, and instance segmentation results outperforming a strong baseline based on optical flow. We show an overview of the proposed method in Figure 7.

7.1.10. Depth-based Hand Pose Estimation: Methods, Data, and Challenges

Participants: James S. Supancic [UC Irvine], Grégory Rogez, Yi Yang [Baidu Research], Jamie Shotton [Microsoft Research], Deva Ramanan [Carnegie Mellon University].

Hand pose estimation has matured rapidly in recent years. The introduction of commodity depth sensors and a multitude of practical applications have spurred new advances. In [16], we provide an extensive analysis of the state-of-the-art, focusing on hand pose estimation from a single depth frame. We summarize important conclusions here: (1) Pose estimation appears roughly solved for scenes with isolated hands. However, methods still struggle to analyze cluttered scenes where hands may be interacting with nearby objects and surfaces. To spur further progress we introduce a challenging new dataset with diverse, cluttered scenes. (2) Many methods evaluate themselves with disparate criteria , making comparisons difficult. We define a

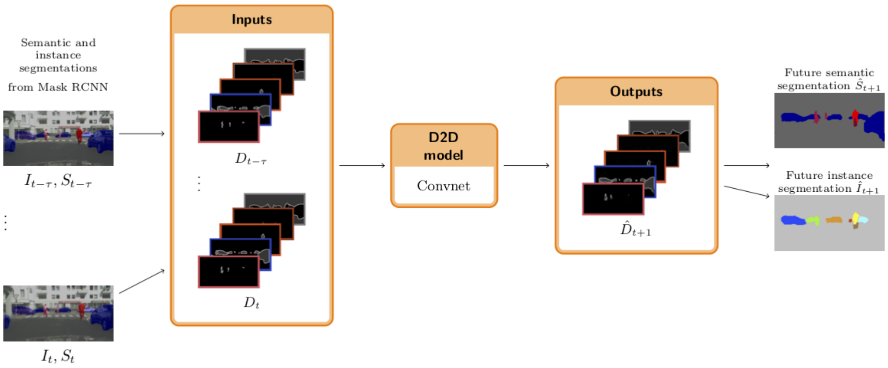

Semantic and instance segmentations from Mask RCNN
$I_{t-\tau}, S_{t-\tau}$
I_t, S_t
Inputs
$D_{t-\tau}$
D_t
D2D model
Convnet
Outputs
$\hat{D}_{t+1}$
Future semantic segmentation $\hat{S}_{t+1}$
Future instance segmentation $\hat{I}_{t+1}$

Figure 7. Our representation enables both future semantic and instance segmentation prediction. It relies on distance maps from the different objects contours: For each channel of an input segmentation, corresponding to a specific class, the segmentation is decomposed into zeros for background, ones for objects and high values for contours. Then a convnet is trained to predict the future representation. Taking its argmax lets us recover the future semantic segmentation, and computing a watershed from it leads to the future instance segmentation.

consistent evaluation criteria, rigorously motivated by human experiments. (3) We introduce a simple nearest-neighbor baseline that outperforms most existing systems (see results in Fig. 8). This implies that most systems do not generalize beyond their training sets. This also reinforces the under-appreciated point that training data is as important as the model itself. We conclude with directions for future progress.

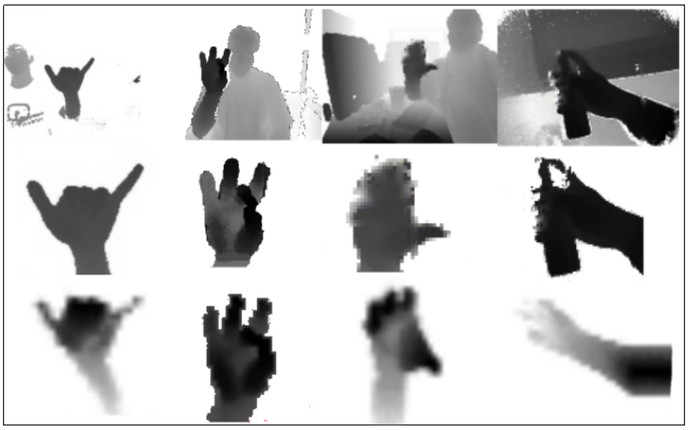


Figure 8. We evaluate a broad collection of hand pose estimation algorithms on different training and testsets under consistent criteria. Test sets which contained limited variety, in pose and range, or which lacked complex backgrounds were notably easier. To aid our analysis, we introduce a simple 3D exemplar (nearest-neighbor) baseline that both detects and estimates pose suprisingly well, outperforming most existing systems. We show the best-matching detection window in (middle) and the best-matching exemplar in (bottom). We use our baseline to rank dataset difficulty, compare algorithms, and show the importance of training set design.

7.1.11. Image-based Synthesis for Deep 3D Human Pose Estimation

Participants: Grégory Rogez, Cordelia Schmid.

In [14], we address the problem of 3D human pose estimation in the wild. A significant challenge is the lack of training data, i.e., 2D images of humans annotated with 3D poses. Such data is necessary to train state-of-the-art CNN architectures. Here, we propose a solution to generate a large set of photorealistic synthetic images of humans with 3D pose annotations. We introduce an image-based synthesis engine that artificially augments a dataset of real images with 2D human pose annotations using 3D Motion Capture (MoCap) data. Given a candidate 3D pose our algorithm selects for each joint an image whose 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. See examples in Figure 9. The resulting images are used to train an end-to-end CNN for full-body 3D pose estimation. We cluster the training data into a large number of pose classes and tackle pose estimation as a K-way classification problem. Such an approach is viable only with large training sets such as ours. Our method outperforms the state of the art in terms of 3D pose estimation in controlled environments (Human3.6M) and shows promising results for in-the-wild images (LSP). This demonstrates that CNNs trained on artificial images generalize well to real images. Compared to data generated from more classical rendering engines, our synthetic images do not require any domain adaptation or fine-tuning stage.

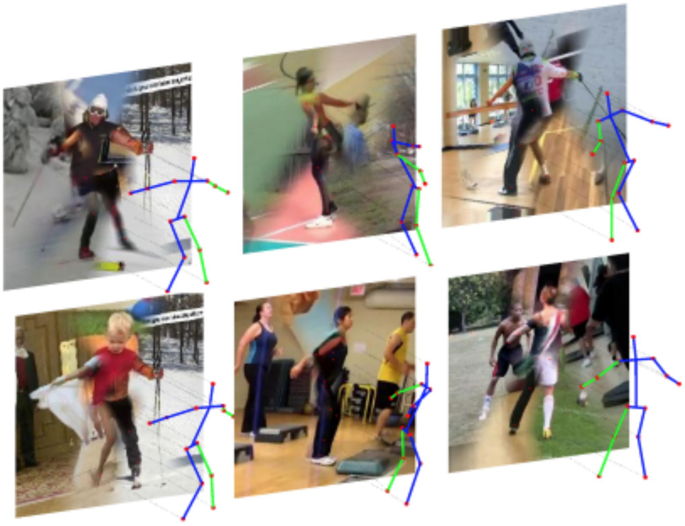


Figure 9. Given a candidate 3D pose, our algorithm selects for each joint an image whose annotated 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. We show 6 examples corresponding to the same 3D pose observed from 6 different camera viewpoints.

7.1.12. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images

Participants: Grégory Rogez, Philippe Weinzaepfel [Naver Labs Europe], Cordelia Schmid.

In [15], 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. See example in Figure 10. Hence, our approach does not require an approximate localization of the humans for initialization. Our architecture, named LCR-Net, 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. All three stages share the convolutional feature layers and are trained jointly. The final pose estimation is obtained by integrating over neighboring pose hypotheses , which is shown to improve over a standard non maximum suppression algorithm. Our approach significantly outperforms the state of the art in 3D pose estimation on Human3.6M, a controlled environment. Moreover, it shows promising results on real images for both single and multi-person subsets of the MPII 2D pose benchmark and demonstrates satisfying 3D pose results even for multi-person images.

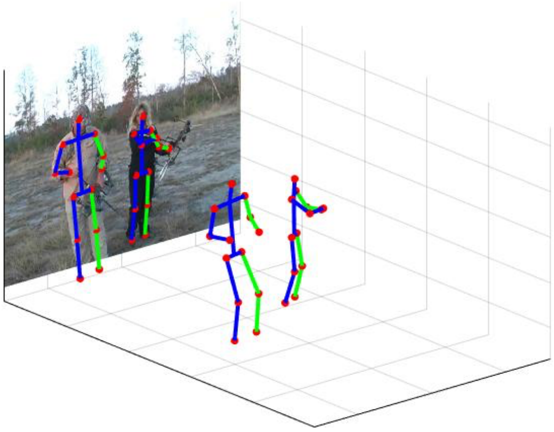


Figure 10. Examples of joint 2D-3D pose detections in a natural image. Even in case of occlusion or truncation, we estimate the joint locations by reasoning in term of full-body 2D-3D poses.

7.1.13. Link and code: Fast indexing with graphs and compact regression codes

Participants: Matthijs Douze [Facebook AI Research], Alexandre Sablayrolles, Hervé Jégou [Facebook AI Research].

Similarity search approaches based on graph walks have recently attained outstanding speed-accuracy trade-offs, taking aside the memory requirements. In [24], we revisit these approaches by considering, additionally, the memory constraint required to index billions of images on a single server. This leads us to propose a method based both on graph traversal and compact representations. We encode the indexed vectors using quantization and exploit the graph structure to refine the similarity estimation, see Figure 11. In essence, our method takes the best of these two worlds: the search strategy is based on nested graphs, thereby providing high precision with a relatively small set of comparisons. At the same time it offers a significant memory compression. As a result, our approach outperforms the state of the art on operating points considering 64–128 bytes per vector, as demonstrated by our results on two billion-scale public benchmarks.

7.1.14. Sparse weakly supervised models for object localization in road environment

Participants: Valentina Zadrija [Univ. Zagreb], Josip Krapac [Univ. Zagreb], Sinisa Segvic [Univ. Zagreb], Jakob Verbeek.

In [19] we propose a novel weakly supervised object localization method based on Fisher-embedding of low-level features (CNN, SIFT), and model sparsity at the component level. Fisher-embedding provides an interesting alternative to raw low-level features, since it allows fast and accurate scoring of image subwindows with a model trained on entire images. Model sparsity reduces overfitting and enables fast evaluation. We also propose two new techniques for improving performance when our method is combined with nonlinear normalizations of the aggregated Fisher representation of the image. These techniques are i) intra-component

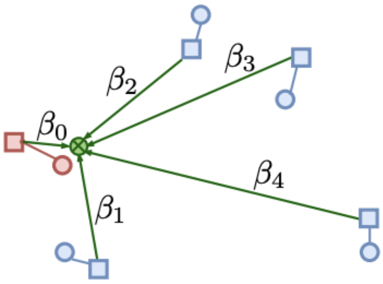


Figure 11. Illustration of our approach: we adopt a graph traversal strategy that maintains a connectivity between all database points. We further improve the estimate by regressing each database vector from its encoded neighbors.

metric normalization and ii) first-order approximation to the score of a normalized image representation. We evaluate our weakly supervised localization method on real traffic scenes acquired from driver's perspective. The method dramatically improves the localization AP over the dense non-normalized Fisher vector baseline (16 percentage points for zebra crossings, 21 percentage points for traffic signs) and leads to a huge gain in execution speed (91× for zebra crossings, 74× for traffic signs). See Figure 12 for several example outputs.

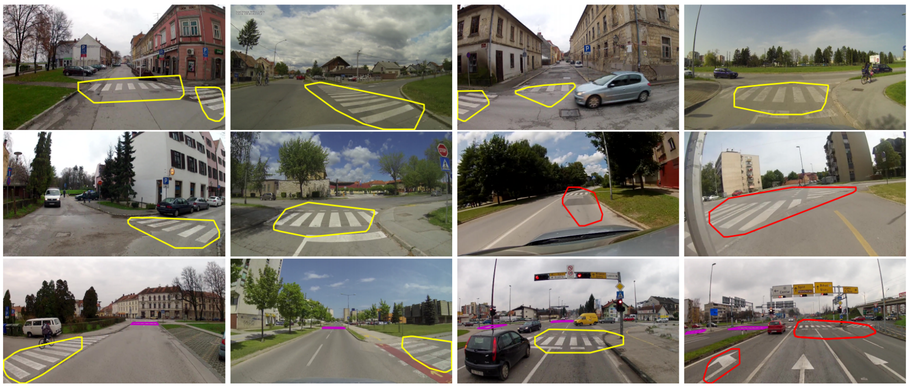


Figure 12. Localization results on test images: correct localization polygons (yellow), false postive responses (red), and ground-truth polygons for false negatives (magenta).

7.1.15. Scene Coordinate Regression with Angle-Based Reprojection Loss for Camera Relocalization

Participants: Xiaotian Li [Aalto Univ.], Juha Ylioinas [Aalto Univ.], Jakob Verbeek, Juho Kannala [Univ. Oulu].

Image-based camera relocalization is an important problem in computer vision and robotics. Recent works utilize convolutional neural networks (CNNs) to regress for pixels in a query image their corresponding 3D world coordinates in the scene. The final pose is then solved via a RANSAC-based optimization scheme using the predicted coordinates, see Figure 13. Usually, the CNN is trained with ground truth scene coordinates, but it has also been shown that the network can discover 3D scene geometry automatically by minimizing single-view reprojection loss. However, due to the deficiencies of reprojection loss, the network needs to be carefully initialized. In [30], we present a new angle-based reprojection loss which resolves the issues of the original reprojection loss. With this new loss function, the network can be trained without careful initialization, and the system achieves more accurate results. The new loss also enables us to utilize available multi-view constraints, which further improve performance.

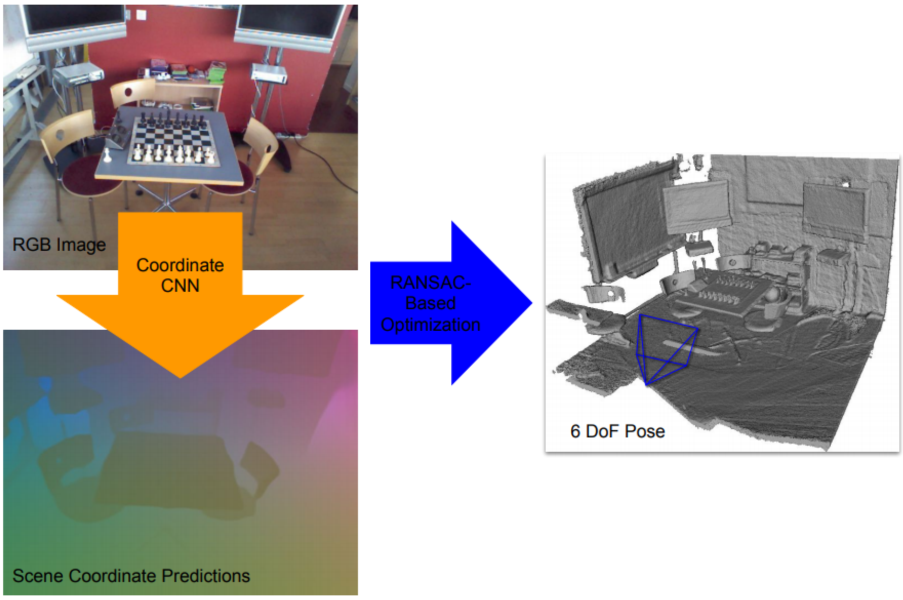


Figure 13. Localization pipeline. In this two-stage pipeline, a coordinate CNN first produces scene coordinate predictions from an RGB image, and then the predicted correspondences are fed into a RANSAC-based solver to determine the camera pose.

7.1.16. FeaStNet: Feature-Steered Graph Convolutions for 3D Shape Analysis

Participants: Nitika Verma, Edmond Boyer [Inria, MORPHEO], Jakob Verbeek.

Convolutional neural networks (CNNs) have massively impacted visual recognition in 2D images, and are now ubiquitous in state-of-the-art approaches. While CNNs naturally extend to other domains, such as audio and video, where data is also organized in rectangular grids, they do not easily generalize to other types of data such as 3D shape meshes, social network graphs or molecular graphs. In our recent paper [38], we propose a novel graph-convolutional network architecture to handle such data. The architecture builds on a generic formulation that relaxes the 1-to-1 correspondence between filter weights and data elements around the center of the convolution, see Figure 14 for an illustration. The main novelty of our architecture is that the shape of the filter is a function of the features in the previous network layer, which is learned as an integral part of the neural network. Experimental evaluations on digit recognition and 3D shape correspondence yield state-of-the-art results, significantly improving over previous work for shape correspondence.

7.2. Statistical Machine Learning

7.2.1. Modulated Policy Hierarchies

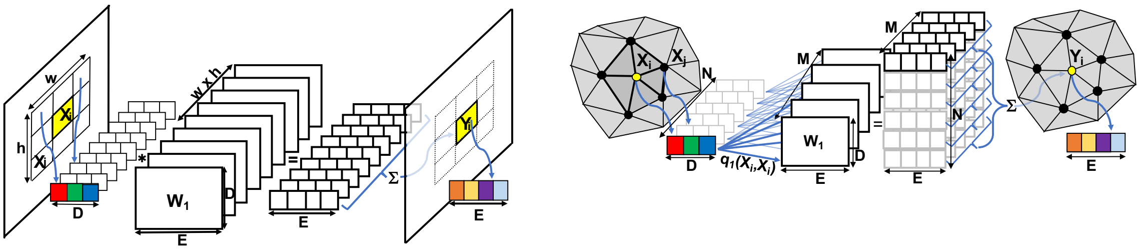
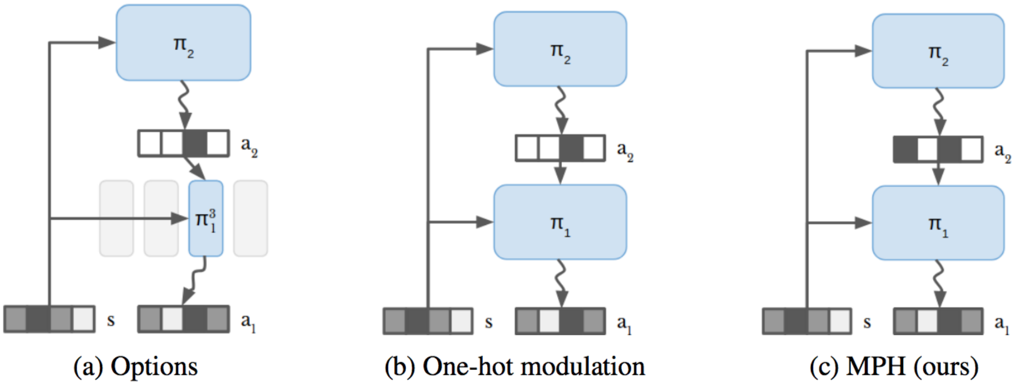


Figure 14. Left: Illustration of a standard CNN, representing the parameters as a set of $M = w \times h$ weight matrices, each of size $D \times E$. Each weight matrix is associated with a single relative position in the input patch. Right: Our graph convolutional network, where each node in the input patch is associated in a soft manner to each of the M weight matrices based on its features using the weight $q_m(\mathbf{x}_i, \mathbf{x}_j)$.

Participants: Alexander Pashevich, Danijar Hafner [Google Brain], James Davidson [Vernalis (R&D) Ltd.], Rahul Sukthankar [Google], Cordelia Schmid.

Solving tasks with sparse rewards is a main challenge in reinforcement learning. While hierarchical controllers are an intuitive approach to this problem, current methods often require manual reward shaping, alternating training phases, or manually defined sub tasks. In [48], we introduce modulated policy hierarchies (MPH), that can learn end-to-end to solve tasks from sparse rewards. To achieve this, we study different modulation signals and exploration for hierarchical controllers. Specifically, we find that communicating via bit-vectors is more efficient than selecting one out of multiple skills, as it enables mixing between them (see Figure 15). To facilitate exploration, MPH uses its different time scales for temporally extended intrinsic motivation at each level of the hierarchy. We evaluate MPH on the robotics tasks of pushing and sparse block stacking, where it outperforms recent baselines.



(a) Options (b) One-hot modulation (c) MPH (ours)

Figure 15. Overview of hierarchical policies. (a) The options agent selects between separate skill networks using a categorical master policy. (b) The one-hot agent combines the skills into a single network and is modulated by a 1-hot signal. (c) Our modulated policy hierarchy sends a binary vector, allowing for richer communication and mixing of skills.

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

Participants: Alberto Bietti, Julien Mairal.

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 [4], 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.

7.2.3. 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 [40], 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.

7.2.4. Learning Disentangled Representations with Reference-Based Variational Autoencoders

Participants: Adria Ruiz, Oriol Martinez [Universitat Pompeu Fabra, Barcelona], Xavier Binefa [Universitat Pompeu Fabra, Barcelona], Jakob Verbeek.

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 [50] 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. 16 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.

7.2.5. On Regularization and Robustness of Deep Neural Networks

Participants: Alberto Bietti, Grégoire Mialon, Julien Mairal.

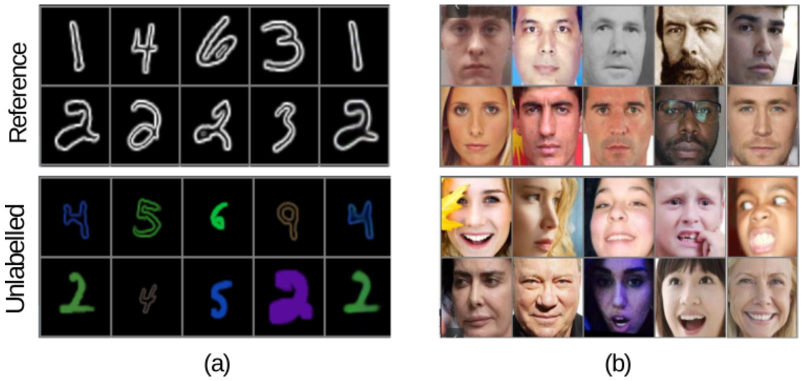


Figure 16. Illustration of different reference-based disentangling problems. (a) Disentangling style from digits. The reference distribution is composed by numbers with a fixed style (b) Disentangling factors of variations related with facial expressions. Reference images correspond to neutral faces. Note that pairing information between unlabelled and reference images is not available during training.

For many supervised learning tasks, deep neural networks are known to work well when large amounts of annotated data are available. Yet, Despite their success, deep neural networks suffer from several drawbacks: they lack robustness to small changes of input data known as “adversarial examples” and training them with small amounts of annotated data is challenging. In [41], we study the connection between regularization and robustness of deep neural networks by viewing them as elements of a reproducing kernel Hilbert space (RKHS) of functions and by regularizing them using the RKHS norm. Even though this norm cannot be computed, we consider various approximations based on upper and lower bounds. These approximations lead to new strategies for regularization, but also to existing ones such as spectral norm penalties or constraints, gradient penalties, or adversarial training. Besides, the kernel framework allows us to obtain margin-based bounds on adversarial generalization. We show that our new algorithms lead to empirical benefits for learning on small datasets and learning adversarially robust models. We also discuss implications of our regularization framework for learning implicit generative models.

7.2.6. Mixed batches and symmetric discriminators for GAN training

Participants: Thomas Lucas, Corentin Tallec [Inria, TAU], Jakob Verbeek, Yann Ollivier [Facebook AI Research].

Generative adversarial networks (GANs) are powerful generative models based on providing feedback to a generative network via a discriminator network. However, the discriminator usually assesses individual samples. This prevents the discriminator from accessing global distributional statistics of generated samples, and often leads to *mode dropping*: the generator models only part of the target distribution. In [32] we propose to feed the discriminator with *mixed batches* of true and fake samples, and train it to predict the ratio of true samples in the batch. The latter score does not depend on the order of samples in a batch. Rather than learning this invariance, we introduce a generic permutation-invariant discriminator architecture, which is illustrated in Figure 17. This architecture is provably a universal approximator of all symmetric functions. Experimentally, our approach reduces mode collapse in GANs on two synthetic datasets, and obtains good results on the CIFAR10 and CelebA datasets, both qualitatively and quantitatively.

7.2.7. Auxiliary Guided Autoregressive Variational Autoencoders

Participants: Thomas Lucas, Jakob Verbeek.

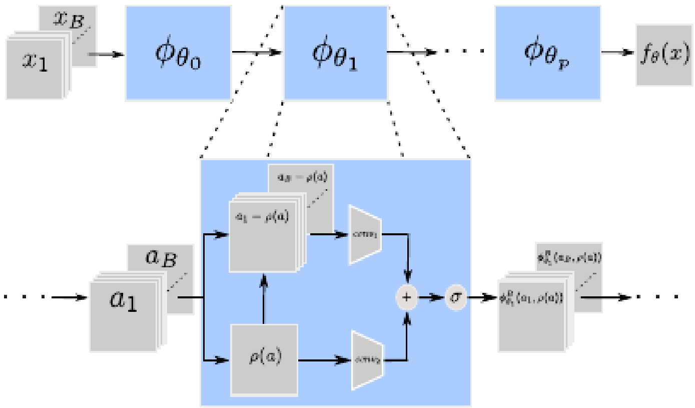


Figure 17. Graphical representation of our discriminator architecture. Each convolutional layer of an otherwise classical CNN architecture is modified to include permutation invariant batch statistics, denoted $\rho(x)$. This is repeated at every layer so that the network gradually builds up more complex statistics.

Generative modeling of high-dimensional data is a key problem in machine learning. Successful approaches include latent variable models and autoregressive models. The complementary strengths of these approaches, to model global and local image statistics respectively, suggest hybrid models combining the strengths of both. Our contribution in [33] is to train such hybrid models using an auxiliary loss function that controls which information is captured by the latent variables and what is left to the autoregressive decoder, as illustrated in Figure 18. In contrast, prior work on such hybrid models needed to limit the capacity of the autoregressive decoder to prevent degenerate models that ignore the latent variables and only rely on autoregressive modeling. Our approach results in models with meaningful latent variable representations, and which rely on powerful autoregressive decoders to model image details. Our model generates qualitatively convincing samples, and yields state-of-the-art quantitative results.

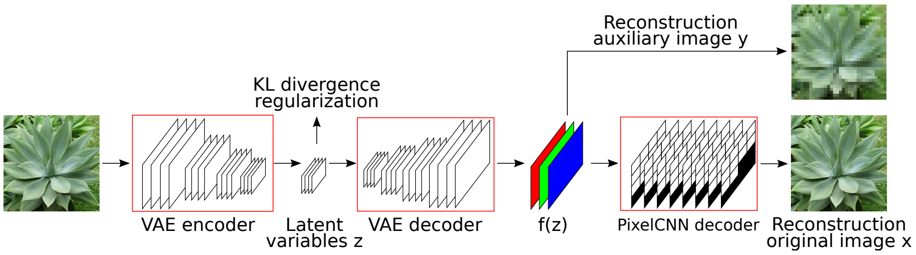


Figure 18. Schematic illustration of our auxiliary guided autoregressive variational autoencoder (AGAVE). An input image is encoded into a latent representation and decoded back into an image. This first reconstruction is guided by an auxiliary maximum likelihood loss and regularized with a Kullback-Liebler divergence. An autoregressive model is then conditionned on the auxiliary reconstruction and also trained with maximum likelihood.

7.2.8. *End-to-End Incremental Learning*

Participants: Francisco Castro [Univ. Malaga], Manuel Marin-Jimenez [Univ. Cordoba], Nicolas Guil [Univ. Malaga], Cordelia Schmid, Karteek Alahari.

Although deep learning approaches have stood out in recent years due to their state-of-the-art results, they continue to suffer from catastrophic forgetting, a dramatic decrease in overall performance when training with new classes added incrementally. This is due to current neural network architectures requiring the entire dataset, consisting of all the samples from the old as well as the new classes, to update the model—a requirement that becomes easily unsustainable as the number of classes grows. We address this issue with our approach [20] to learn deep neural networks incrementally, using new data and only a small exemplar set corresponding to samples from the old classes. This is based on a loss composed of a distillation measure to retain the knowledge acquired from the old classes, and a cross-entropy loss to learn the new classes. Our incremental training is achieved while keeping the entire framework end-to-end, i.e., learning the data representation and the classifier jointly, unlike recent methods with no such guarantees. We evaluate our method extensively on the CIFAR-100 and ImageNet (ILSVRC 2012) image classification datasets, and show state-of-the-art performance.

7.3. Large-scale Optimization for Machine Learning

7.3.1. *Stochastic Subsampling for Factorizing Huge Matrices*

Participants: Julien Mairal, Arthur Mensch [Inria, Parietal], Gael Varoquaux [Inria, Parietal], Bertrand Thirion [Inria, Parietal].

In [13], we present a matrix-factorization algorithm that scales to input matrices with both huge number of rows and columns. Learned factors may be sparse or dense and/or non-negative, which makes our algorithm suitable for dictionary learning, sparse component analysis, and non-negative matrix factorization. Our algorithm streams matrix columns while subsampling them to iteratively learn the matrix factors. At each iteration, the row dimension of a new sample is reduced by subsampling, resulting in lower time complexity compared to a simple streaming algorithm. Our method comes with convergence guarantees to reach a stationary point of the matrix-factorization problem. We demonstrate its efficiency on massive functional Magnetic Resonance Imaging data (2 TB), and on patches extracted from hyperspectral images (103 GB). For both problems, which involve different penalties on rows and columns, we obtain significant speed-ups compared to state-of-the-art algorithms. The main principle of the method is illustrated in Figure 19.

7.3.2. *An Inexact Variable Metric Proximal Point Algorithm for Generic Quasi-Newton Acceleration*

Participants: Hongzhou Lin, Julien Mairal, Zaid Harchaoui [Univ. Washington].

In [12], we propose a generic approach to accelerate gradient-based optimization algorithms with quasi-Newton principles. The proposed scheme, called QuickeNing, can be applied to incremental first-order methods such as stochastic variance-reduced gradient (SVRG) or incremental surrogate optimization (MISO). It is also compatible with composite objectives, meaning that it has the ability to provide exactly sparse solutions when the objective involves a sparsity-inducing regularization. QuickeNing relies on limited-memory BFGS rules, making it appropriate for solving high-dimensional optimization problems. Besides, it enjoys a worst-case linear convergence rate for strongly convex problems. We present experimental results where QuickeNing gives significant improvements over competing methods for solving large-scale high-dimensional machine learning problems, see Figure 20 for example.

7.3.3. *Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice*

Participants: Hongzhou Lin, Julien Mairal, Zaid Harchaoui [Univ. Washington].

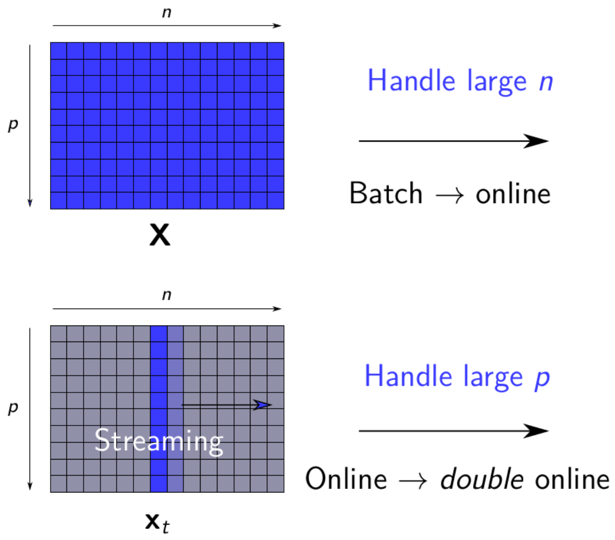


Figure 19. Illustration of the matrix factorization algorithm, which streams columns in one dimension while subsampling them.

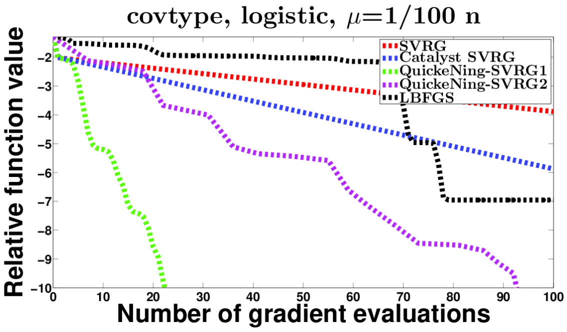


Figure 20. An illustration of the minimization of logistic regression. Significant improvement is observed by applying QuickeNing.

In [11], we introduce a generic scheme for accelerating gradient-based optimization methods in the sense of Nesterov. The approach, called Catalyst, builds upon the inexact accelerated proximal point algorithm for minimizing a convex objective function, and consists of approximately solving a sequence of well-chosen auxiliary problems, leading to faster convergence. One of the key to achieve acceleration in theory and in practice is to solve these sub-problems with appropriate accuracy by using the right stopping criterion and the right warm-start strategy. In this work, we give practical guidelines to use Catalyst and present a comprehensive theoretical analysis of its global complexity. We show that Catalyst applies to a large class of algorithms, including gradient descent, block coordinate descent, incremental algorithms such as SAG, SAGA, SDCA, SVRG, Finito/MISO, and their proximal variants. For all of these methods, we provide acceleration and explicit support for non-strongly convex objectives. We conclude with extensive experiments showing that acceleration is useful in practice, especially for ill-conditioned problems.

7.3.4. Catalyst Acceleration for Gradient-Based Non-Convex Optimization

Participants: Courtney Paquette [Univ. Washington], Hongzhou Lin, Dmitriy Drusvyatskiy [Univ. Washington], Julien Mairal, Zaid Harchaoui [Univ. Washington].

In [34], we introduce a generic scheme to solve nonconvex optimization problems using gradient-based algorithms originally designed for minimizing convex functions. When the objective is convex, the proposed approach enjoys the same properties as the Catalyst approach of Lin et al, 2015. When the objective is nonconvex, it achieves the best known convergence rate to stationary points for first-order methods. Specifically, the proposed algorithm does not require knowledge about the convexity of the objective; yet, it obtains an overall worst-case efficiency of $O(\epsilon^{-2})$ and, if the function is convex, the complexity reduces to the near-optimal rate $O(\epsilon^{-2/3})$. We conclude the paper by showing promising experimental results obtained by applying the proposed approach to SVRG and SAGA for sparse matrix factorization and for learning neural networks (see Figure 21).

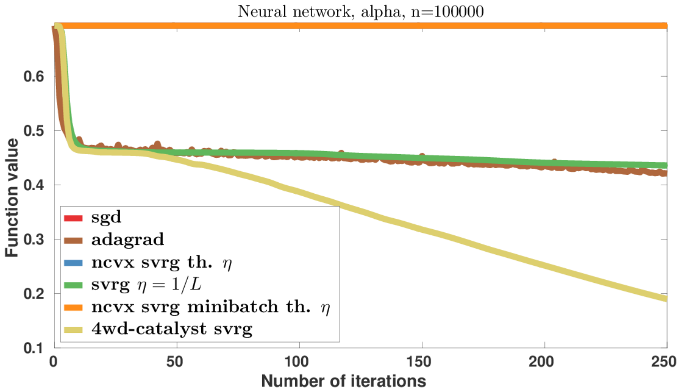


Figure 21. Comparison of different algorithms for the minimization of a two-layer neural network. Applying our method provides a clear acceleration in terms of function value.

7.4. Pluri-disciplinary Research

7.4.1. Biological Sequence Modeling with Convolutional Kernel Networks

Participants: Dexiong Chen, Laurent Jacob [CNRS, LBBE Laboratory], 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 [5], we introduce a hybrid approach between convolutional neural networks and kernel methods to model biological sequences. Our method 22 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.

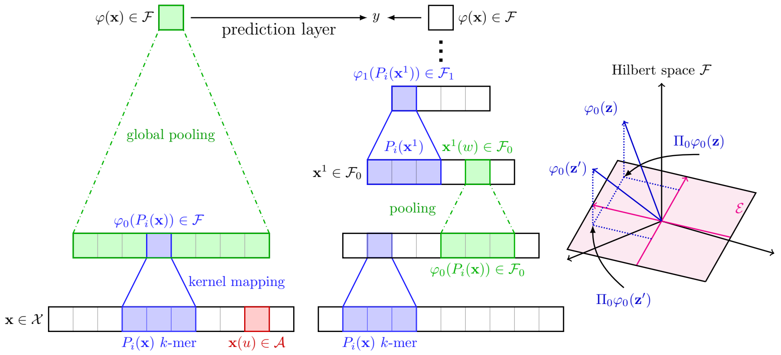

$\varphi(\mathbf{x}) \in \mathcal{F}$
prediction layer
y
$\varphi(\mathbf{x}) \in \mathcal{F}$
global pooling
$\varphi_0(P_i(\mathbf{x})) \in \mathcal{F}$
kernel mapping
$\mathbf{x} \in \mathcal{X}$
$P_i(\mathbf{x})$ k-mer
$\mathbf{x}(u) \in \mathcal{A}$
$\varphi_1(P_i(\mathbf{x}^1)) \in \mathcal{F}_1$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

Figure 22. 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 $\mathfrak{F}$ 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.

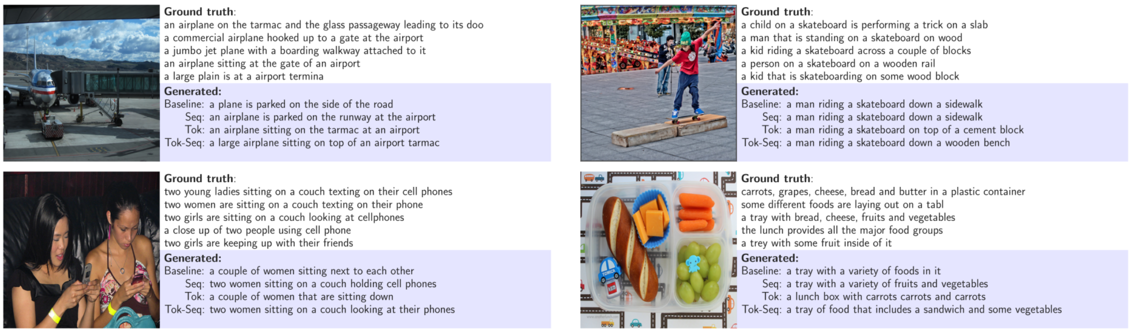
7.4.2. Token-level and sequence-level loss smoothing for RNN language models

Participants: Maha Elbayad, Laurent Besacier [LIG], Jakob Verbeek.

In [28] we investigate the limitations of the maximum likelihood estimation (MLE) used when training recurrent neural network language models. First, the MLE treats all sentences that do not match the ground truth as equally poor, ignoring the structure of the output space. Second, it suffers from "exposure bias": during training tokens are predicted given ground-truth sequences, while at test time prediction is conditioned on generated output sequences. To overcome these limitations we build upon the recent reward augmented maximum likelihood approach i.e., sequence-level smoothing that encourages the model to predict sentences close to the ground truth according to a given performance metric. We extend this approach to token-level loss smoothing, and propose improvements to the sequence-level smoothing approach. Our experiments on two different tasks, image captioning (see Fig. 23) and machine translation, show that token-level and sequence-level loss smoothing are complementary, and significantly improve results.

7.4.3. Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction

Participants: Maha Elbayad, Laurent Besacier [LIG], Jakob Verbeek.



Ground truth:
an airplane on the tarmac and the glass passageway leading to its doo
a commercial airplane hooked up to a gate at the airport
a jumbo jet plane with a boarding walkway attached to it
an airplane sitting at the gate of an airport
a large plain is at a airport termina

Generated:
Baseline: a plane is parked on the side of the road
Seq: an airplane is parked on the runway at the airport
Tok: an airplane sitting on the tarmac at an airport
Tok-Seq: a large airplane sitting on top of an airport tarmac

Ground truth:
a child on a skateboard is performing a trick on a slab
a man that is standing on a skateboard on wood
a kid riding a skateboard across a couple of blocks
a person on a skateboard on a wooden rail
a kid that is skateboarding on some wood block

Generated:
Baseline: a man riding a skateboard down a sidewalk
Seq: a man riding a skateboard down a sidewalk
Tok: a man riding a skateboard on top of a cement block
Tok-Seq: a man riding a skateboard down a wooden bench

Ground truth:
two young ladies sitting on a couch texting on their cell phones
two women are sitting on a couch texting on their phone
two girls are sitting on a couch looking at cellphones
a close up of two people using cell phone
two girls are keeping up with their friends

Generated:
Baseline: a couple of women sitting next to each other
Seq: two women sitting on a couch holding cell phones
Tok: a couple of women that are sitting down
Tok-Seq: two women sitting on a couch looking at their phones

Ground truth:
carrots, grapes, cheese, bread and butter in a plastic container
some different foods are laying out on a tabl
a tray with bread, cheese, fruits and vegetables
the lunch provides all the major food groups
a trey with some fruit inside of it

Generated:
Baseline: a tray with a variety of foods in it
Seq: a tray with a variety of fruits and vegetables
Tok: a lunch box with carrots carrots and carrots
Tok-Seq: a tray of food that includes a sandwich and some vegetables

Figure 23. Examples of generated captions with the baseline MLE and our models with attention.

Current state-of-the-art machine translation systems are based on encoder-decoder architectures, that first encode the input sequence, and then generate an output sequence based on the input encoding. Both are interfaced with an attention mechanism that recombines a fixed encoding of the source tokens based on the decoder state. In [27], we propose an alternative approach which instead relies on a single 2D convolutional neural network across both sequences as illustrated in Figure 24. Each layer of our network re-codes source tokens on the basis of the output sequence produced so far. Attention-like properties are therefore pervasive throughout the network. Our model yields excellent results, outperforming state-of-the-art encoder-decoder systems, while being conceptually simpler and having fewer parameters.

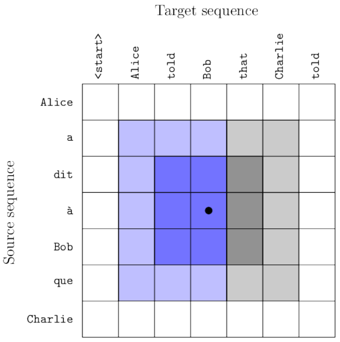


Figure 24. Convolutional layers in our model use masked 3×3 filters so that features are only computed from previous output symbols. Illustration of the receptive fields after one (dark blue) and two layers (light blue), together with the masked part of the field of view of a normal 3×3 filter (gray)

7.4.4. Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis

Participant: Ghislain Durif.

The development of high-throughput biology technologies now allows the investigation of the genome-wide diversity of transcription in single cells. This diversity has shown two faces: the expression dynamics (gene to gene variability) can be quantified more accurately, thanks to the measurement of lowly-expressed genes. Second, the cell-to-cell variability is high, with a low proportion of cells expressing the same gene at the same time/level. Those emerging patterns appear to be very challenging from the statistical point of view, especially to represent and to provide a summarized view of single-cell expression data. PCA is one of the most powerful framework to provide a suitable representation of high dimensional datasets, by searching for latent directions catching the most variability in the data. Unfortunately, classical PCA is based on Euclidean distances and projections that work poorly in presence of over-dispersed counts that show drop-out events (zero-inflation) like single-cell expression data. In [25], we propose a probabilistic Count Matrix Factorization (pCMF) approach for single-cell expression data analysis, that relies on a sparse Gamma-Poisson factor model. This hierarchical model is inferred using a variational EM algorithm. We show how this probabilistic framework induces a geometry that is suitable for single-cell data visualization, and produces a compression of the data that is very powerful for clustering purposes. Our method is competed to other standard representation methods like t-SNE, and we illustrate its performance for the representation of zero-inflated over-dispersed count data. We also illustrate our work with results on a publicly available data set, being single-cell expression profile of neural stem cells. Our work is implemented in the pCMF R-package.

7.4.5. Extracting Universal Representations of Cognition across Brain-Imaging Studies

Participants: Arthur Mensch [Inria, Parietal], Julien Mairal, Bertrand Thirion [Inria, Parietal], Gael Varoquaux [Inria, Parietal].

We show in [47] how to extract shared brain representations that predict mental processes across many cognitive neuroimaging studies. Focused cognitive-neuroimaging experiments study precise mental processes with carefully-designed cognitive paradigms; however the cost of imaging limits their statistical power. On the other hand, large-scale databasing efforts increase considerably the sample sizes, but cannot ask precise cognitive questions. To address this tension, we develop new methods that turn the heterogeneous cognitive information held in different task-fMRI studies into common-universal-cognitive models. Our approach does not assume any prior knowledge of the commonalities shared by the studies in the corpus; those are inferred during model training. The method uses deep-learning techniques to extract representations - task-optimized networks - that form a set of basis cognitive dimensions relevant to the psychological manipulations, as illustrated in Figure 25. In this sense, it forms a novel kind of functional atlas, optimized to capture mental state across many functional-imaging experiments. As it bridges information on the neural support of mental processes, this representation improves decoding performance for 80% of the 35 widely-different functional imaging studies that we consider. Our approach opens new ways of extracting information from brain maps, increasing statistical power even for focused cognitive neuroimaging studies, in particular for those with few subjects.

7.4.6. Loter: Inferring local ancestry for a wide range of species

Participants: Thomas Dias-Alves, Julien Mairal, Michael Blum [CNRS, TIMC Laboratory].

Admixture between populations provides opportunity to study biological adaptation and phenotypic variation. Admixture studies can rely on local ancestry inference for admixed individuals, which consists of computing at each locus the number of copies that originate from ancestral source populations, as illustrated in Figure 26. Existing software packages for local ancestry inference are tuned to provide accurate results on human data and recent admixture events. In [7], we introduce Loter, an open-source software package that does not require any biological parameter besides haplotype data in order to make local ancestry inference available for a wide range of species. Using simulations, we compare the performance of Loter to HAPMIX, LAMP-LD, and RFMix. HAPMIX is the only software severely impacted by imperfect haplotype reconstruction. Loter is the less impacted software by increasing admixture time when considering simulated and admixed human genotypes. LAMP-LD and RFMIX are the most accurate method when admixture took place 20 generations ago or less; Loter accuracy is comparable or better than RFMix accuracy when admixture took place of 50 or more generations; and its accuracy is the largest when admixture is more ancient than 150 generations. For simulations of admixed Populus genotypes, Loter and LAMP-LD are robust to increasing admixture times by contrast to RFMix. When comparing length of reconstructed and true ancestry tracts, Loter and LAMP-LD provide results whose accuracy is again more robust than RFMix to increasing admixture times. We apply Loter to admixed Populus individuals and lengths of ancestry tracts indicate that admixture took place around

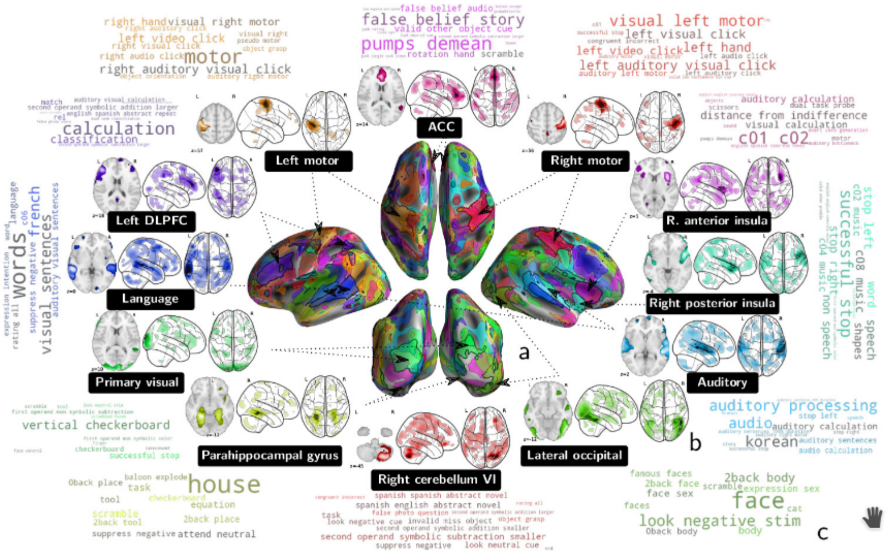
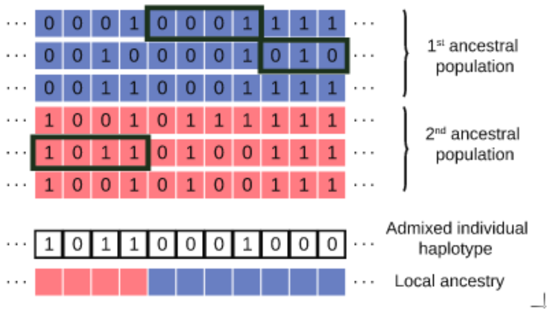


Figure 25. Visualization of some of task-optimized networks. Our approach allows to learn networks that are important for inter-subject decoding across studies. These networks, individually focal and collectively well spread across the cortex, are readily associated with the cognitive tasks that they contribute to predict. We display a selection of these networks, named with the salient anatomical brain region they recruit, along with a word-cloud representation of the stimuli whose likelihood increases with the network activation.

100 generations ago. The Loter software package and its source code are available at https://github.com/bcm-uga/Loter.



··· 0 0 0 1 0 0 0 1 1 1 1 ···
··· 0 0 1 0 0 0 0 1 0 1 0 ···
··· 0 0 1 1 0 0 0 1 1 1 1 ···
··· 1 0 0 1 0 1 1 1 1 1 1 ···
··· 1 0 1 1 0 1 0 0 1 1 1 ···
··· 1 0 0 1 0 1 0 0 1 1 1 ···

··· 1 0 1 1 0 0 0 1 0 0 0 ···

Figure 26. Graphical description of Local Ancestry Inference as implemented in the software Loter. Given a collection of parental haplotypes from the source populations depicted in blue and red, Loter assumes that an haplotype of an admixed individuals is modeled as a mosaic of existing parental haplotypes.

8. Bilateral Contracts and Grants with Industry

8.1. MSR-Inria joint lab: structured large-scale machine learning

Participants: Julien Mairal, Alberto Bietti.

Machine learning is now ubiquitous in industry, science, engineering, and personal life. While early successes were obtained by applying off-the-shelf techniques, there are two main challeges faced by machine learning in the “ big data ” era : structure and scale. The project proposes to explore three axes, from theoretical, algorithmic and practical perspectives: (1) large-scale convex optimization, (2) large-scale combinatorial optimization and (3) sequential decision making for structured data. The project involves two Inria sites and four MSR sites and started at the end of 2013. Alberto Bietti visited MSR New York in 2018.

8.2. Amazon

Participants: Grégory Rogez, Cordelia Schmid.

We received an Amazon Faculty Research Award in 2018. The objective is 3D human action recognition from monocular RGB videos. The idea is to extend our recent work on human 3D pose estimation to videos and to develop an approach for action recognition based on temporal pose based on appropriate 3D features.

8.3. 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.

8.4. Facebook

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

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.

8.5. NAVER LABS Europe

Participants: Cordelia Schmid, Karteek Alahari, Julien Mairal, Jakob Verbeek, Vasileios Choutas, Nieves Crasto.

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.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. 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).

9.1.2. 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.

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, Thomas Dias-Alves, 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. 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 new 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, Pavel Tokmakov, Konstantin Shmelkov, Vladyslav Sydorov, Daan Wynen, Mikita 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) - Robotics Institute - Deva Ramanan

Start year: 2016

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

We propose to form an associate team GAYA, with the primary goal of interpreting videos in terms of recognizing actions, understanding the human-human and human-object interactions. Despite several years of research, it is yet unclear what is an efficient and robust video representation to attack this challenge. In order to address this, GAYA will focus on building semantic models, wherein we learn the video feature representation 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 (LEAR and WILLOW) and a US university (Carnegie Mellon University [CMU]). It will allow the three teams to effectively combine their respective strengths in areas such as inference and machine learning approaches for vision tasks, feature representation, large-scale learning, geometric reasoning. The main expected outcomes of this collaboration are: effective learnt representations of video content, new machine learning algorithms for handling minimally annotated data, large-scale public datasets for benchmarking, theoretical analysis of objects shapes and contours.

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. She spent one month at MPI in April 2018. In 2018, the project resulted in the development of an approach for object interaction.
- **University of Washington:** Julien Mairal collaborates with Zaid Harchaoui, former member of the team, on the topic of large-scale optimization.

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 from Sep to Dec 2018.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

- C. Schmid is one of the general chairs for ECCV 2020.

10.1.1.2. Member of the Organizing Committees

- Several permanent members of the team co-organized the international summer school PAISS 2018.
- J. Mairal is a member of the organizing committee for the international conference SIAM Imaging Science 2020.
- J. Mairal co-organized the Journées SMAI-MODE, which will took place in March 2018.
- J. Mairal is a co-organizer of the workshop OSL'19 at Les Houches.
- G. Rogez was one of the organizers of the CVPR workshop on Human Pose, Motion, Activities and Shape in 3D (3D HUMANS 2018).
- C. Schmid was one of the organizers of Workshop on Artificial Intelligence, Horizon Maths, Paris, 2018.

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- K. Alahari: Area chair for CVPR 2018, ICCV 2019.
- K. Alahari: Senior program committee member for IJCAI 2018, IJCAI 2019.
- J. Mairal: Area chair for ICML and NeurIPS 2018.
- C. Schmid: Area chair for NeurIPS 2018, ICML 2018, ECCV 2018, ICCV 2019.
- J. Verbeek: Area chair for ECCV 2018, ICCV 2019.

10.1.2.2. Reviewer

The permanent members of the team reviewed numerous papers for numerous international conferences in computer vision and machine learning, including CVPR, ECCV, NeurIPS, ICML.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- K. Alahari: Associate editor for Computer Vision and Image Understanding journal, since 2018.
- 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.
- C. Schmid: Editor in Chief of the International Journal of Computer Vision, since 2013.
- C. Schmid: Associate editor for Foundations and Trends in Computer Graphics and Vision, since 2005.
- J. Verbeek: Associate editor for Image and Vision Computing Journal, 2011-2018.
- J. Verbeek: Associate editor for International Journal on Computer Vision, since 2014.
- J. Verbeek: Associate editor for IEEE Transactions Pattern Analysis and Machine Intelligence, since 2018.

10.1.3.2. Reviewer - Reviewing Activities

The permanent members of the team reviewed numerous papers for numerous international journals in computer vision (IJCV, PAMI, CVIU), machine learning (JMLR, Machine Learning). Some of them also review for journals in optimization (SIAM Journal on Optimization, Mathematical Programming), image processing (SIAM Imaging Science).

10.1.4. Invited Talks

- K. Alahari: Invited talk at IISc Bangalore, India, Dec 2018.
- K. Alahari: Invited talk at Hyderabad AI Symposium, India, Dec 2018.
- K. Alahari: Invited seminar at Simon Fraser Univ., Vancouver, Canada, Nov 2018.
- K. Alahari: Invited seminar at POSTECH, Pohang, South Korea, Oct 2018.
- K. Alahari: COVIEW workshop, ACM Multimedia, Seoul, South Korea, Oct 2018.
- K. Alahari: Le Futur des Images workshop, IXXI, Lyon, Oct 2018.
- K. Alahari: Invited seminar at Universidad de Malaga, Apr 2018.
- K. Alahari: CVPR AC Meeting, Toronto, Feb 2018.
- A. Bietti: Séminaire de statistiques et machine learning, Telecom ParisTech.
- A. Bietti: Laplace reading group, ENS Paris.
- A. Bietti: Grenoble Optimization Days, LJK, UGA.
- G. Durif: Seminar at LJK, UGA, May 2018.
- G. Durif: Seminar at Institut de Recherche Mathematique Avancée (IRMA), Strasbourg Univ., June 2018.
- M. Elbayad: Seminar at NAVER LABS Europe, Grenoble, Sep 2018.
- P. Luc: Invited talk in the class “Deep Learning for Image Analysis” at Ecole des Mines.
- J. Mairal: Seminaire Parisien d’Optimisation, Paris.
- J. Mairal: Workshop on the Future of Random Projections, Paris.
- J. Mairal: talk in mini-symposium at ISMP, Bordeaux.
- J. Mairal: keynote in Theory of Deep Learning Workshop, ICML 2018, Stockholm.
- J. Mairal: keynote in CEFRL workshop, ECCV 2018, Munich.
- J. Mairal: invited speaker at AI and ML workshop, Telecom ParisTech.
- J. Mairal: Seminar, Gatsby Unit, UCL, London.

- J. Mairal: Seminar, College de France, Paris.
- J. Mairal: Seminar, University of Geneva.
- J. Mairal: Seminar, Université d'Avignon.
- J. Mairal: Seminar, Université de Marseille.
- G. Rogez: Invited speaker, Apple Seminar, Salt Lake City, June 2018.
- G. Rogez: Seminar, NAVER LABS Europe, Grenoble, December 2018.
- C. Schmid: Invited speaker at the Google Multimodal Machine Perception Workshop, San Francisco, October 2018.
- C. Schmid: Invited speaker, “What is optical flow for?” workshop at ECCV, September 2018.
- C. Schmid: Invited speaker, 3rd Intl. Workshop on Video Segmentation at ECCV, September 2018.
- C. Schmid: Keynote speaker at Deep Learning Conference, Rennes, September 2018.
- C. Schmid: Keynote speaker at ActivityNet workshop, in conjunction with CVPR, June 2018.
- C. Schmid: Invited talk at CVPR Good Citizen of CVPR event, in conjunction with CVPR, June 2018.
- C. Schmid: Keynote speaker at 3D Humans workkshop, in conjunction with CVPR, June 2018.
- C. Schmid: Invited speaker at Symposium on AI, Académie des sciences, Paris, February 2018.
- C. Schmid: Presentation at Prairie/industry meeting, Paris, December 2018.
- C. Schmid: Presentation at Google workshop on 3D Deep Learning, October 2018.
- C. Schmid: Seminar for AI residents, Google Mountain View, June 2018.
- C. Schmid: Seminar at MPI Tübingen, April 2018.
- C. Schmid: Seminars at Google Zürich (April), Mountain View and Paris (March).
- C. Schmid: Seminar at Leopoldina section meeting, Ulm, February 2018.
- J. Verbeek: Seminar Facebook AI Research, Paris, December 2018.
- J. Verbeek: Seminar NAVER LABS Europe, Grenoble, December 2018.
- J. Verbeek: Seminar Univ. Amsterdam, July 2018.
- J. Verbeek: Workshop Mathematics and Deep Learning at Univ. Aix-Marseille, November 2018.
- J. Verbeek: ARC6 Deep Learning & Deep Reinforcement Learning Workshop, Lyon.

10.1.5. Scientific Expertise

- J. Mairal: Reviewer for ERC (Consolidator and Starting).
- J. Mairal: Panel member for ANR.
- J. Mairal: Judge for the IBM Watson AI Xprize.
- G. Rogez: reviewer for ANR.

10.1.6. Research Administration

- J. Mairal: Participation in the setting up of the 3IA institute in Grenoble.
- C. Schmid: Member, board of directors of the Computer Vision Foundation (CVF), since 2016.
- C. Schmid: Member, PAMI-TC awards committee and the PAMI-TC executive commitee.
- J. Verbeek: Member, 2018 recruitment committee for an Assistant Professor position at Univ. Grenoble Alpes, Laboratoire Jean Kuntzmann (LJK).
- J. Verbeek: Member, steering committee MinaLogic, innovation cluster for digital technologies based in France's Auvergne-Rhône-Alpes region, since 2018.

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: A. Bietti, Mini-course “Optimization for large-scale machine learning”, 10h eqTD, conference SMAI-MODE, Autrans, France.

Doctorat: J. Mairal, Mini-course “Optimization for large-scale machine learning”, 10h eqTD, conference SMAI-MODE, Autrans, France.

Doctorat: J. Mairal, Mini-course “Optimization for large-scale machine learning”, 4.5h eqTD, conference Mascot-Num, Nantes, France.

Doctorat: J. Mairal, Lecturer at the YSU - ISTC Joint Summer School on Machine Learning, 12h eqTD, Yerevan, Armenia.

Doctorat: J. Mairal, Lecturer at the PAISS summer school, 2h eqTD, Grenoble, France.

Doctorat: C. Schmid, Course on action recognition, 2h eqTD, PRAIRIE Artifical Intelligence Summer School, July 2018, Grenoble, France.

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: C. Schmid, Object recognition and computer vision, 9h eqTD, M2, ENS Paris, France

Master: J. Verbeek, K. Alahari, C. Schmid, Machine Learning and Object Recognition, 27h eqTD, M2, Grenoble University, France

Master: J. Verbeek and J. Mairal, Advanced Learning Models, 27h eqTD, M2, UGA, Grenoble.

10.2.2. Supervision

PhD: N. Chesneau, Learning to Recognize Actions with Weak Supervision, Univ. Grenoble Alpes, Feb 2018, Karteek Alahari and Cordelia Schmid.

PhD: A. Mensch, Apprentissage de représentations en imagerie fonctionnelle, Univ. Paris-Saclay, Sep 2018, Gael Varoquaux, Bertrand Thirion and Julien Mairal.

PhD: P. Tokmakov, Apprentissage à partir du mouvement, Univ. Grenoble Alpes, Jun 2018, Karteek Alahari and Cordelia Schmid.

10.2.3. Juries

K. Alahari: Mostafa S. Ibrahim, November 2018, external examiner, Simon Fraser University.

K. Alahari: Ignacio Rocco Spremolla, 2018, member of “comité de suivi de thèse”, ENS Paris.

J. Mairal: Magda Gregorova, 2018, rapporteur, University of Geneva.

J. Mairal: Saeed Varasteh, December 2018, examinateur, Univ. Grenoble Alpes.

J. Mairal: Olga Permiakova, November 2018, member of “comité de suivi de thèse”, Univ. Grenoble Alpes.

J. Mairal: Vincent Prost, November 2018, member of “comité de suivi de thèse”, Univ. Paris Saclay.

G. Rogez: Francisco Castro, December 2018, jury member, Univ. Malaga.

C. Schmid: Jean-Baptise Alayrac, September 2018, examinateur, ENS Paris.

C. Schmid: Stéphane Lathuilière, May 2018, presidente, UGA.

C. Schmid: Gunnar Atli Sigurdsson, May 2018, thesis proposal, CMU.

C. Schmid: Christoph Lassner, April 2018, rapporteur, Universitaet Tuebingen.

C. Schmid: Maxime Oquab, January 2018, examinatuer, ENS Paris.

J. Verbeek: Riccardo Del Chiaro, 2018-2020, member of supervisory committee, Univ. Florence.

J. Verbeek: Fabien Baradel, 2017-2019, member of supervisory committee, INSA Lyon.

J. Verbeek: Mélanie Ducoffe, 2018, rapporteur, Univ. Côte d'Azur, Sophia-Antipolis.

10.3. Popularization

- P. Luc: Interviews with "Computer Vision News" and "Les Echos Start".
- J. Mairal: intervention grand public sur l'IA lors d'un événement organisé par PWN (Professional Woman Network) à Lyon en octobre 2018.
- J. Mairal: conférence grand public sur l'IA avec Cédric Villani en juin 2018, mairie de Lyon.
- J. Mairal: participation à un débat public sur l'IA organisé par Mme la député Cendra Motin en avril 2018, Montalieu-Vercieu.

10.3.1. Internal or external Inria responsibilities

- C. Schmid: Member, "Comité scientifique", Inria Grenoble, since 2015.
- J. Verbeek: Scientific correspondent, national project calls, Inria Grenoble, since 2017.
- J. Verbeek: Member, Inria Grenoble working group on HPC - Big Data - Machine learning, since 2018.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] N. CHESNEAU. *Learning to Recognize Actions with Weak Supervision*, Université Grenoble Alpes, February 2018, https://tel.archives-ouvertes.fr/tel-01893147

[2] G. CHÉRON. *Structured modeling and recognition of human actions in video*, Ecole normale supérieure - ENS PARIS, December 2018, https://hal.inria.fr/tel-01975247

[3] P. TOKMAKOV. *Learning from motion*, Université Grenoble Alpes, June 2018, https://tel.archives-ouvertes.fr/tel-01908817

Articles in International Peer-Reviewed Journals

[4] 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} 25, pp. 1-49, https://arxiv.org/abs/1706.03078 , https://hal.inria.fr/hal-01536004

[5] D. CHEN, L. JACOB, J. MAIRAL. *Biological Sequence Modeling with Convolutional Kernel Networks*, in "Bioinformatics", February 2019 [*DOI :* 10.1093/BIOINFORMATICS/BTZ094], https://hal.inria.fr/hal-01632912

[6] N. Chesneau, K. Alahari, C. Schmid. *Learning from Web Videos for Event Classification*, in "IEEE Transactions on Circuits and Systems for Video Technology", October 2018, vol. 28, n^o 10, pp. 3019-3029 [*DOI :* 10.1109/TCSVT.2017.2764624], https://hal.inria.fr/hal-01618400

[7] T. Dias-Alves, J. Mairal, M. Blum. *Loter: A software package to infer local ancestry for a wide range of species*, in "Molecular Biology and Evolution", June 2018, vol. 35, n^o 9, pp. 2318 - 2326 [*DOI :* 10.1093/molbev/msy126], https://hal.inria.fr/hal-01630228

[8] G. Durif, L. Modolo, J. Michaelsson, J. E. Mold, S. Lambert-Lacroix, F. Picard. *High Dimensional Classification with combined Adaptive Sparse PLS and Logistic Regression*, in "Bioinformatics", February 2018, vol. 34, n^o 3, pp. 485-493, https://arxiv.org/abs/1502.05933 [*DOI :* 10.1093/bioinformatics/btx571], https://hal.archives-ouvertes.fr/hal-01587360

[9] B. Ham, M. Cho, C. Schmid, J. Ponce. *Proposal Flow: Semantic Correspondences from Object Proposals*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", July 2018, vol. 40, n^o 7, pp. 1711-1725 [*DOI :* 10.1109/TPAMI.2017.2724510], https://hal.inria.fr/hal-01644132

[10] G. Hu, X. Peng, Y. Yang, T. Hospedales, J. Verbeek. *Frankenstein: Learning Deep Face Representations using Small Data*, in "IEEE Transactions on Image Processing", January 2018, vol. 27, n^o 1, pp. 293-303, https://arxiv.org/abs/1603.06470 [*DOI :* 10.1109/TIP.2017.2756450], https://hal.inria.fr/hal-01306168

[11] H. Lin, J. Mairal, Z. Harchaoui. *Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice*, in "Journal of Machine Learning Research", April 2018, vol. 18, n^o 212, pp. 1-54, http://jmlr.org/papers/volume18/17-748/17-748.pdf, https://hal.inria.fr/hal-01664934

[12] H. Lin, J. Mairal, Z. Harchaoui. *An Inexact Variable Metric Proximal Point Algorithm for Generic Quasi-Newton Acceleration*, in "SIAM Journal on Optimization", 2019, https://arxiv.org/abs/1610.00960 , forthcoming, https://hal.inria.fr/hal-01376079

[13] A. Mensch, J. Mairal, B. Thirion, G. Varoquaux. *Stochastic Subsampling for Factorizing Huge Matrices*, in "IEEE Transactions on Signal Processing", January 2018, vol. 66, n^o 1, pp. 113-128, https://arxiv.org/abs/1701.05363 [*DOI :* 10.1109/TSP.2017.2752697], https://hal.archives-ouvertes.fr/hal-01431618

[14] G. Rogez, C. Schmid. *Image-based Synthesis for Deep 3D Human Pose Estimation*, in "International Journal of Computer Vision", September 2018, vol. 126, n^o 9, pp. 993–1008, https://arxiv.org/abs/1802.04216 [*DOI :* 10.1007/s11263-018-1071-9], https://hal.inria.fr/hal-01717188

[15] 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

[16] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, D. Ramanan. *Depth-based hand pose estimation: methods, data, and challenges*, in "International Journal of Computer Vision", November 2018, vol. 126, n^o 11, pp. 1180–1198 [*DOI :* 10.1007/s11263-018-1081-7], https://hal.inria.fr/hal-01759416

[17] 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

[18] G. VAROL, I. LAPTEV, C. SCHMID. *Long-term Temporal Convolutions for Action Recognition*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", June 2018, vol. 40, n^{o} 6, pp. 1510-1517, https://arxiv.org/abs/1604.04494 [*DOI :* 10.1109/TPAMI.2017.2712608], https://hal.inria.fr/hal-01241518

[19] V. ZADRIJA, J. KRAPAC, S. ŠEGVIĆ, J. VERBEEK. *Sparse weakly supervised models for object localization in road environment*, in "Computer Vision and Image Understanding", November 2018, vol. 176-177, pp. 9-21 [*DOI :* 10.1016/J.CVIU.2018.10.004], https://hal.inria.fr/hal-01900418

International Conferences with Proceedings

[20] F. M. CASTRO, M. J. MARÍN-JIMÉNEZ, N. GUIL, C. SCHMID, K. ALAHARI. *End-to-End Incremental Learning*, in "ECCV 2018 - European Conference on Computer Vision", Munich, Germany, V. FERRARI, M. HEBERT, C. SMINCHISESCU, Y. WEISS (editors), Lecture Notes in Computer Science, Springer, September 2018, vol. 11216, pp. 241-257, https://arxiv.org/abs/1807.09536 [*DOI :* 10.1007/978-3-030-01258-8_15], https://hal.inria.fr/hal-01849366

[21] V. CHOUTAS, P. WEINZAEPFEL, J. REVAUD, C. SCHMID. *PoTion: Pose MoTion Representation for Action Recognition*, in "CVPR 2018 - IEEE Conference on Computer Vision and Pattern Recognition", Salt Lake City, United States, IEEE, June 2018, pp. 7024-7033 [*DOI :* 10.1109/CVPR.2018.00734], https://hal.inria.fr/hal-01764222

[22] G. CHÉRON, J.-B. ALAYRAC, I. LAPTEV, C. SCHMID. *A flexible model for training action localization with varying levels of supervision*, in "NIPS 2018 - 32nd Conference on Neural Information Processing Systems", Montréal, Canada, December 2018, pp. 1-17, https://arxiv.org/abs/1806.11328 , https://hal.inria.fr/hal-01937002

[23] C. COUPRIE, P. LUC, J. VERBEEK. *Joint Future Semantic and Instance Segmentation Prediction*, in "ECCV Workshop on Anticipating Human Behavior", Munich, Germany, Lecture Notes in Computer Science, Springer, September 2018, vol. 11131, pp. 154-168 [*DOI :* 10.1007/978-3-030-11015-4_14], https://hal.inria.fr/hal-01867746

[24] M. DOUZE, A. SABLAYROLLES, H. JÉGOU. *Link and code: Fast indexing with graphs and compact regression codes*, in "CVPR 2018 - IEEE Conference on Computer Vision & Pattern Recognition", Salt Lake City, United States, IEEE, June 2018, pp. 1-9, https://hal.inria.fr/hal-01955971

[25] G. DURIF, L. MODOLO, J. E. MOLD, S. LAMBERT-LACROIX, F. PICARD. *Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis*, in "RECOMB 2018 - 22nd Annual International Conference on Research in Computational Molecular Biology", Paris, France, B. J. RAPHAEL (editor), Lecture Notes in Bioinformatics, Springer, April 2018, vol. 10812, pp. 254-255, https://hal.archives-ouvertes.fr/hal-01962030

[26] N. DVORNIK, J. MAIRAL, C. SCHMID. *Modeling Visual Context is Key to Augmenting Object Detection Datasets*, in "ECCV 2018 - European Conference on Computer Vision", Munich, Germany, Lecture Notes in Computer Science, Springer, September 2018, vol. 11216, pp. 375-391, https://arxiv.org/abs/1807.07428 [*DOI :* 10.1007/978-3-030-01258-8_23], https://hal.archives-ouvertes.fr/hal-01844474

[27] M. ELBAYAD, L. BESACIER, J. VERBEEK. *Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction*, in "CoNLL 2018 - Conference on Computational Natural Language Learning", Brussels, Belgium, ACL, October 2018, pp. 97–107, https://hal.inria.fr/hal-01851612

[28] M. Elbayad, L. Besacier, J. Verbeek. *Token-level and sequence-level loss smoothing for RNN language models*, in "ACL - 56th Annual Meeting of the Association for Computational Linguistics", Melbourne, Australia, ACL, July 2018, pp. 2094–2103, https://hal.inria.fr/hal-01790879

[29] C. Gu, C. Sun, D. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. R. Sukthankar, C. Schmid, J. Malik. *AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions*, in "CVPR 2018 - Computer Vision and Pattern Recognition", Salt Lake City, United States, IEEE, June 2018, pp. 6047-6056 [*DOI :* 10.1109/CVPR.2018.00633], https://hal.inria.fr/hal-01764300

[30] X. Li, J. Ylioinas, J. Verbeek, J. Kannala. *Scene Coordinate Regression with Angle-Based Reprojection Loss for Camera Relocalization*, in "ECCV 2018 - Workshop Geometry Meets Deep Learning", Munich, Germany, Lecture Notes in Computer Science, Springer, September 2018, vol. 11131, pp. 229-245 [*DOI :* 10.1007/978-3-030-11015-4_19], https://hal.inria.fr/hal-01867143

[31] P. Luc, C. Couprie, Y. Lecun, J. Verbeek. *Predicting Future Instance Segmentation by Forecasting Convolutional Features*, in "ECCV 2018 - European Conference on Computer Vision", Munich, Germany, Lecture Notes in Computer Science, Springer, September 2018, vol. 11213, pp. 593-608 [*DOI :* 10.1007/978-3-030-01240-3_36], https://hal.inria.fr/hal-01757669

[32] T. Lucas, C. Tallec, J. Verbeek, Y. Ollivier. *Mixed batches and symmetric discriminators for GAN training*, in "ICML - 35th International Conference on Machine Learning", Stockholm, Sweden, Proceedings of Machine Learning Research, July 2018, vol. 80, pp. 2844-2853, https://hal.inria.fr/hal-01791126

[33] T. Lucas, J. Verbeek. *Auxiliary Guided Autoregressive Variational Autoencoders*, in "ECML-PKDD 2018 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Dublin, Ireland, Lecture Notes in Computer Science, Springer, September 2018, vol. 11051, pp. 443-458 [*DOI :* 10.1007/978-3-030-10925-7_27], https://hal.inria.fr/hal-01652881

[34] C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, Z. Harchaoui. *Catalyst for Gradient-based Nonconvex Optimization*, in "AISTATS 2018 - 21st International Conference on Artificial Intelligence and Statistics", Lanzarote, Spain, Proceedings of Machine Learning Research, April 2018, vol. 84, pp. 613-622, https://hal.inria.fr/hal-01773296

[35] K. Shmelkov, C. Schmid, K. Alahari. *How good is my GAN?*, in "ECCV 2018 - European Conference on Computer Vision", Munich, Germany, V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (editors), Lecture Notes in Computer Science, Springer, September 2018, vol. 11206, pp. 218-234, https://arxiv.org/abs/1807.09499 [*DOI :* 10.1007/978-3-030-01216-8_14], https://hal.inria.fr/hal-01850447

[36] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, K. Alahari. *Actor and Observer: Joint Modeling of First and Third-Person Videos*, in "CVPR 2018 - IEEE Conference on Computer Vision & Pattern Recognition", Salt Lake City, Utah, United States, IEEE, June 2018, pp. 7396-7404, https://arxiv.org/abs/1804.09627 [*DOI :* 10.1109/CVPR.2018.00772], https://hal.inria.fr/hal-01755547

[37] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, C. Schmid. *BodyNet: Volumetric Inference of 3D Human Body Shapes*, in "ECCV 2018 - 15th European Conference on Computer Vision", Munich, Germany, Lecture Notes in Computer Science, Springer, September 2018, vol. 11211, pp. 20-38, https://arxiv.org/abs/1804.04875 [*DOI :* 10.1007/978-3-030-01234-2_2], https://hal.inria.fr/hal-01852169

[38] N. VERMA, E. BOYER, J. VERBEEK. *FeaStNet: Feature-Steered Graph Convolutions for 3D Shape Analysis*, in "CVPR - IEEE Conference on Computer Vision & Pattern Recognition", Salt Lake City, United States, IEEE, June 2018, pp. 2598-2606, https://arxiv.org/abs/1706.05206 [*DOI :* 10.1109/CVPR.2018.00275], https://hal.inria.fr/hal-01540389

[39] D. WYNEN, C. SCHMID, J. MAIRAL. *Unsupervised Learning of Artistic Styles with Archetypal Style Analysis*, in "NeurIPS 2018 - Annual Conference on Neural Information Processing Systems", Montréal, Canada, December 2018, pp. 1-10, Accepted at NIPS 2018, Montréal, Canada, https://hal.inria.fr/hal-01802131

Other Publications

[40] A. BIETTI, A. AGARWAL, J. LANGFORD. *A Contextual Bandit Bake-off*, December 2018, https://arxiv.org/abs/1802.04064 - working paper or preprint, https://hal.inria.fr/hal-01708310

[41] A. BIETTI, G. MIALON, D. CHEN, J. MAIRAL. *A Kernel Perspective for Regularizing Deep Neural Networks*, January 2019, https://arxiv.org/abs/1810.00363 - working paper or preprint, https://hal.inria.fr/hal-01884632

[42] 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

[43] N. DVORNIK, J. MAIRAL, C. SCHMID. *On the Importance of Visual Context for Data Augmentation in Scene Understanding*, September 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01869784

[44] 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

[45] T. LUCAS, K. SHMELKOV, K. ALAHARI, C. SCHMID, J. VERBEEK. *Adversarial training of partially invertible variational autoencoders*, February 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02007787

[46] T. LUCAS, K. SHMELKOV, K. ALAHARI, C. SCHMID, J. VERBEEK. *Adversarial training of partially invertible variational autoencoders*, March 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01886285

[47] A. MENSCH, J. MAIRAL, B. THIRION, G. VAROQUAUX. *Extracting Universal Representations of Cognition across Brain-Imaging Studies*, October 2018, https://arxiv.org/abs/1809.06035 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01874713

[48] A. PASHEVICH, D. HAFNER, J. DAVIDSON, R. R. SUKTHANKAR, C. SCHMID. *Modulated Policy Hierarchies*, December 2018, Deep RL workshop at NIPS 2018, https://hal.archives-ouvertes.fr/hal-01963580

[49] J. PEYRE, I. LAPTEV, C. SCHMID, J. SIVIC. *Detecting rare visual relations using analogies*, January 2019, https://arxiv.org/abs/1812.05736 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01975760

[50] A. RUIZ, O. MARTINEZ, X. BINEFA, J. VERBEEK. *Learning Disentangled Representations with Reference-Based Variational Autoencoders*, January 2019, working paper or preprint, https://hal.inria.fr/hal-01896007