AIMH Research Activities 2020

Nicola Aloia, Giuseppe Amato, Valentina Bartalesi, Filippo Benedetti, Paolo Bolettieri, Fabio Carrara, Vittore Casarosa, Luca Ciampi, Cesare Concordia, Silvia Corbara, Marco Di Benedetto, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, Gabriele Lagani, Fabio Valerio Massoli, Carlo Meghini, Nicola Messina, Daniele Metilli, Alessio Molinari, Alejandro Moreo, Alessandro Nardi, Andrea Pedrotti, Nicolò Pratelli, Fausto Rabitti, Pasquale Savino, Fabrizio Sebastiani, Costantino Thanos, Luca Trupiano, Lucia Vadicamo, Claudio Vairo

Abstract
The Artificial Intelligence for Media and Humanities laboratory (AIMH) has the mission to investigate and advance the state of the art in the Artificial Intelligence field, specifically addressing applications to digital media and digital humanities, and taking also into account issues related to scalability. This report summarize the 2020 activities of the research group.

Keywords
Multimedia Information Retrieval – Artificial Intelligence — Computer Vision — Similarity Search – Machine Learning for Text – Text Classification – Transfer learning – Representation Learning

[1] *AIMH Lab, ISTI-CNR, via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy*
*__Corresponding author__: giuseppe.amato@isti.cnr.it

Contents

AIMH
ARTIFICIAL INTELLIGENCE FOR MEDIA AND HUMANITIES

http://aimh.isti.cnr.it

Introduction

The Artificial Intelligence for Media and Humanities laboratory (AIMH) of the Information Science and Technologies Institute "A. Faedo" (ISTI) of the Italian National Research Council (CNR) located in Pisa, has the mission to investigate and advance the state of the art in the Artificial Intelligence field, specifically addressing applications to digital media and digital humanities, and taking also into account issues related to scalability.

The laboratory is composed of four research groups:

AI4Text: The AI4Text is active in the area at the crossroads of machine learning and text analysis; it investigates novel algorithms and methodologies, and novel applications of these to different realms of text analysis. The above-mentioned area includes tasks such as representation learning

for text classification, transfer learning for cross-lingual and cross-domain text classification, sentiment classification, sequence learning for information extraction, text quantification, transductive text classification, cost-sensitive text classification, and applications of the above to domains such as authorship analysis and technology-assisted review. The group consists of Fabrizio Sebastiani (Director of Research), Andrea Esuli (Senior Researcher), Alejandro Moreo (Researcher), Silvia Corbara, Alessio Molinari, and Andrea Pedrotti (PhD Students), and is led by Fabrizio Sebastiani.

Humanities: Investigating AI-based solutions to represent, access, archive, and manage tangible and intangible cultural heritage data. This includes solutions based on ontologies, with a special focus on narratives, and solutions based on multimedia content analysis, recognition, and retrieval. The group consists of Carlo Meghini (Senior Researcher), Valentina Bartalesi, Cesare Concordia (Researchers), Luca Trupiano (Technologist), Daniele Metilli (PhD Student), Filippo Benedetti, Nicolò Pratelli (Graduate Fellows), and Costantino Thanos, Vittore Casarosa, Nicola Aloia (Research Associates), and is led by Carlo Meghini.

Large-scale IR: Investigating efficient, effective, and scalable AI-based solutions for searching multimedia content in large datasets of non-annotated data. This includes techniques for multimedia content extraction and representation, scalable access methods for similarity search, multimedia database management. The group consists of Claudio Gennaro, Pasquale Savino (Senior Researchers), Lucia Vadicamo, Claudio Vairo (Researchers), Paolo Bolettieri (Technician), Luca Ciampi, Gabriele Lagani (PhD Students), and Fausto Rabitti (Research Associate), and is led by Claudio Gennaro.

Multimedia: Investigating new AI-based solutions to image and video content analysis, understanding, and classification. This includes techniques for detection, recognition (object, pedestrian, face, etc), classification, feature extraction (low- and high-level, relational, cross-media, etc), anomaly detection also considering adversarial machine learning threats. The group consists of Giuseppe Amato (Senior Researcher), Fabrizio Falchi, Marco Di Benedetto, Claudio Vairo (Researchers), Alessandro Nardi (Technician), Fabio Carrara, Fabio Massoli (Post-doc Fellows), Nicola Messina (PhD Student), and is led by Fabrizio Falchi.

In this paper, we report the activities of the AIMH research group on 2020. The rest of the report is organized as follows. In Section 1, we summarize the research conducted on our main research fields. In Section 2, we describe the projects in which we were involved during the year. We report the complete list of papers we published in 2020, together with their abstract, in Section 3. The list of Thesis on which we were involved can be found in Section 5. In Section 7, we highlight the datasets we created and made publicly available during 2020.

1. Research Topics

In the following, we report a list of active research topics and subtopics at AIMH in 2020.

1.1 Artificial Intelligence

1.1.1 Adversarial Machine Learning

Adversarial machine learning is about attempting to fool models through malicious input. The topic has become very popular with the recent advances on Deep Learning. We studied this topic with a focus on detection of adversarial examples and images in particular, contributing the field with several publications in the last years. Our research investigated adversarial detection powered by the analysis of the internal activation of deep networks (a.k.a. deep features) collecting encouraging results over the last three years. This year's research activity on the topic included a) the evaluation of adversarial robustness of novel deep architectures [12] (see Section 1.1.4) and b) the analysis of adversarial defenses in a popular safety-critical application — face recognition. For the latter, we considered the task of detecting adversarial faces, i.e., malicious faces given to machine learning systems in order to fool the recognition and the verification evaluations [32, 24]. We considered a critical scenario which is typically met when dealing with real-world face images, i.e., the resolution variations. Thus, considering input data from heterogeneous sources, we studied the dependence of adversarial attacks from the image resolution [31] and obtained interesting insights from the perspective of both, the attacker and the defender.

1.1.2 Deep Anomaly Detection

Anomalies are met in every scientific field. The term "anomaly" is itself source of ambiguity since it is usually used to point at both outliers and anomalies. Training deep learning architectures on the task of detecting such events is challenging since they are rarely observed. Moreover, typically we don't know their origin and the construction of a detaset containing such a kind of data is too expensive. For such a reason, unsupervied and semi-supervised techniques are exploited to train neural networks. The deep one-class classification approach was proposed in 2018, and it appeared to be a promising line of research. In such context, we contributed with two novel methodologies to tackle one-class anomaly detection. In [13], we proposed a deep generative model that combines and generalizes both GANs and AutoEncoders; the former provides realism to the reconstructions, and the latter provides consistency with respect to the input. Both these aspects improved reconstruction-based anomaly detection, in which we spot anomalies by comparing inputs and their reconstructions made by the model. Instead, in [35], we proposed a novel method named MOCCA, in which we exploit the piece-wise nature of deep learning models to detect anomalies. We tasked the model to minimize the deep features distance among a reference point, the class centroid for anomaly-free images, and the current input. By extracting the deep representations

at different depths and combining them, MOCCA improved upon the state-of-the-art considering the one-class classification setting on the the task of anomaly detection.

1.1.3 Hebbian Learning

Traditional neural networks are trained using gradient descent methods with error backpropagation. Despite the great success of such training algorithms, the neuroscientific community has doubts about the biological plausibility of backpropagation learning schemes, proposing a different learning model known as *Hebbian principle*: "Neurons that fire together wire together". Starting from this simple principle, different Hebbian learning variants have been formulated. These approaches are interesting also from a computer science point of view, because they allow to perform common data analysis operations - such as clustering, Principal Component Analysis (PCA), Independent Component Analysis (ICA), and others - in an online, efficient, and neurally plausible fashion. Taking inspiration from biology, we investigate how Hebbian approaches can be integrated with today's machine learning techniques [28], in order to improve the training process in terms of speed, generalization capabilities, sample efficiency.

An even more biologically plausible model of neural computation is based on Spiking Neural Networks (SNNs). In this model, neurons communicate via short pulses called *spikes*. This communication approach is the key towards energy efficiency in the brain. We are using SNNs to accurately simulate real neuronal cultures, in collaboration with neuroscience colleagues, who can produce such cultures in lab. Multi-Electrode Array (MEA) devices can be used to stimulate and record activity from cultured networks, raising the question of whether such cultures can be trained to perform AI tasks. Our simulations help us understand the optimal parameters a cultured network should have in order to solve a given task, providing insights to guide neuroscientists in the creation of real cultures with the desired properties [29].

1.1.4 Neural Ordinary Differential Equations

Presented in a NeurIPS 2018 best paper [16], Neural Ordinary Differential Equations comprise novel differentiable and learnable models (also referred to as ODE-Nets) whose outputs are defined as the solution of a system of parametric ordinary differential equations. Those models exhibit benefits such as a O(1)-memory consumption and a straight-forward modelling of continuous-time and inhomogeneous data, and when using adaptive ODE solvers, they acquire also other interesting properties, such as input-dependent adaptive computation and the tunability (via a tolerance parameter) of the accuracy-speed trade-off at inference time. Neural ODE provides a natural way to model irregular time series in which time between observations brings information and need to be considered. This paves the way for enhanced models in many applications, i.e. model health statuses given irregular medical reports.

During this year, we started experimenting with this new architecture by a) testing its ability to efficiently create flexible image representations [14], and b) measuring its robustness to adversarial perturbations in light of its adaptability properties [12].

1.2 AI and Digital Humanities

The AI & DH group at AIMH employs AI-based methods to research, design and experimentally develop innovative tools to support the work of the scholar humanist. These methods hinge on formal ontologies as powerful tools for the design and the implementation of information systems that exhibit intelligent behavior. Formal ontologies are also regarded as the ideal place where computer scientists and humanists can meet and collaborate to co-create innovative applications that can effectively support the work of the latter. The group pursues in particular the notion of formal narrative as a powerful addition to the information space of digital libraries; an ontology for formal narratives has been developed in the last few years and it is currently being enriched through the research carried out by the members of the group and tested through the validation carried out in the context of the Mingei project. The group is also engaged in the formal representation of literary texts and of the surrounding knowledge, through the HDN project which continues the seminal work that led to the DanteSources application where an ontology-based approach was firstly employed. Finally, through the participation to the ARIADNEplus and the SSHOC projects, the group is actively involved in the making of two fundamental infrastructures in the European landscape, on archaeology and on social sciences humanities, respectively.

1.3 Artificial Intelligence for Text

1.3.1 Learning to quantify

Learning to quantify has to do with training a predictor that estimates the prevalence values of the classes of interest in sample of unlabelled data. This problem has particular relevance in scenarios characterized by distribution shift (which may itself be caused by either covariate shift or prior probability shift), since standard learning algorithms for training classifiers are based on the IID assumption, which is violated in scenarios characterized by distribution drift. The AI4Text group has carried out active research on learning to quantify since 2010.

One of our recent activities in this direction has involved the study of evaluation measures for the learning to quantify task [49]. Such a study is necessary because there is no widespread agreement in the literature on which evaluation measures are the best, or the most appropriate, for evaluating quantification algorithms. Our study has been of the "axiomatic" type, i.e., it has consisted of laying down a number of properties that an evaluation measure for (single-label multiclass) quantification might or might not satisfy, and of proving, for each such property and for each evaluation measure considered, whether the measure satisfies the property or not. The study has brought about some surprising results, which indicate that some measures that were once considered "standard" should instead be deprecated.

Another activity [22] has involved the study of transfer learning techniques for learning to quantify in cross-lingual scenarios, i.e., in situations in which there are little or no training examples for the ("target") language for which we want perform quantification, so that we may want to leverage the training data that we have for some other ("source") language. Our solution involves the combination of a highly successful transfer learning method (Distributional Correspondence Indexing) with a quantification technique based on deep learning (QuaNet), and is the first published solution for the problem of cross-lingual quantification.

In parallel, we have also looked back at past approaches to learning to quantify with a critical eye. In one such study [45] we have reassessed the true merits of "classify and count", the baseline of all quantification studies, due to the fact that, as we have found out, in many published studies this method has been a strawman rather than a baseline, due to lack of or suboptimal parameter optimization. In another such study [44] we have looked back at past research on sentiment quantification, and found that the different approaches to such a task have been compared inappropriately, due to a faulty experimental protocol. We have thus carried out a complete reassessment of these approaches, this time using a much more robust protocol which involves a much more extensive experimentation; our results have upturned past conclusions concerning the relative merits of such approaches.

1.3.2 Learning to classify text

The supervised approach to text classification (TC) is almost 30 years old; despite this, text classification continues to be an active research topic, due to its central role in a number of text analysis and text management tasks.

One aspect we have been working on is feature weighting for TC. The introduction of "supervised term weighting" (STW) more than 15 years ago seemed to establish a landmark; however, STW has since failed to deliver consistent results. Our recent study [41] has investigated, and shown the superiority of, "learning to weight" techniques that allow learning the optimal STW technique from data, i.e., choosing the STW technique that is optimal for our data.

In a different study [43] we have tackled the problem of *cross-lingual TC*, i.e., the task of leveraging training data for a "source" language in order to perform TC in a different, "target" language for which we have little or no training data. In [43] we have extended a previously proposed method for heterogeneous transfer learning (called "Funnelling") to leverage correlations in data that are informative for the TC process; while Funnelling exploit class-class correlations, our "Generalized Funnelling" system also exploits word-class correlations (for which we designed a new type of embedding) and word-word correlations.

1.3.3 Technology-assisted review

Technology-assisted review (TAR) is the task of supporting the work of human annotators who need to "review" automatically labelled data items, i.e., check the correctness of the labels assigned to these items by automatic classifiers. Since only a subset of such items can be feasibly reviewed, the goal of these algorithms is to exactly identify the items whose review is expected to be cost-effective. We have been working on this task since 2018, proposing TAR *risk minimization* algorithms that attempt to strike an optimal tradeoff between the contrasting goals of minimizing the cost of human intervention and maximizing the accuracy of the resulting labelled data. An aspect of TAR we have worked on more recently is improving the quality of the posterior probabilities that the risk minimization algorithm receives as input by an automated classifier. To this end, we have carried out a thorough study of SLD, an algorithm that, while being the state of the art in this task, had insufficiently been studied. Our study [21] has determined exactly in what conditions SLD can be expected to improve the quality of the posterior probabilities (and hence to be beneficial to the downstream TAR algorithm), and has determined that in other conditions SLD can instead bring about a deterioration of this quality.

1.3.4 Authorship analysis

Authorship analysis has to do with training predictors that infer characteristics of the author of a document of unknown paternity. We have worked on a subproblem of authorship analysis called *authorship verification*, which consists in training a binary classifier that decides whether a text of disputed paternity is by a candidate author or not. Specifically, we have concentrated on a renowned case study, the so-called *Epistle to Cangrande*, written in medieval Latin apparently by Dante Alighieri, but whose authenticity has been disputed by scholars in the last century. To this end, we have built and made available to the scientific community two datasets of Medieval Latin texts [20], which we have used for training two separate predictors, one for the first part of the *Epistle* (which has a dedicatory nature) and one for the second part (which is instead a literary essay). The authorship verifiers that we have built indicate, although with different degrees of certainty, that neither the first nor the second part of the *Epistle* are by Dante. These predictions are corroborated by the fact that, once tested according to a leave-one-out experimental protocol on the two datasets, the two predictors exhibit extremely high accuracy [19].

1.4 Artificial Intelligence for Mobility Analysis

1.4.1 Language modeling applied to trajectory classification

Mobility information collected by Location-Based Social Networks (e.g., Foursquare) allow to model mobility at a more abstract and semantically rich level than simple geographical traces. These trace are called multiple-aspect trajectories, and include high level concepts, e.g., going to a theater and then to an Italian restaurant, in addition to the geographical locations. Multiple-aspect trajectories enable to implement new services that exploit similarity models among users based on these high level concepts, rather than simple match of geographical locations. In this context a collaboration among AIMH,

contributing the expertise on language modeling methods, colleagues of the High Performing Computing laboratory, and colleagues of the Universidade Federal de Santa Catarina (Florianópolis, Brazil) led to the development of a novel method for semantic trajectory modeling and classification, Multiple-Aspect tRajectory Classifier (MARC) [47]. MARC uses a trajectory embeddings method derived from the Word2Vec model [40] and then a recurrent neural network to recognize the user who generated it, achieving state-of-the-art results.

1.5 Computer Vision

1.5.1 Learning in Virtual Worlds

In the new spring of artificial intelligence, and in particular in its sub-field known as machine learning, a significant series of important results have shifted the focus of industrial and research communities toward the generation of valuable data from which learning algorithms can be trained. For several applications, in the era of big data, the availability of real input examples, to train machine learning algorithms, is not considered an issue. However, for several other applications, there is not such an abundance of training data. Sometimes, even if data is available it must be manually revised to make it usable as training data (e.g., by adding annotations, class labels, or visual masks), with a considerable cost. In fact, although a series of annotated datasets are available and successfully used to produce important academic results and commercially fruitful products, there is still a huge amount of scenarios where laborious human intervention is needed to produce high quality training sets. For example, such cases include, but are not limited to, safety equipment detection, weapon-wielding detection, and autonomous driven cars.

To overcome these limitations and to provide useful examples in a variety of scenarios, the research community has recently started to leverage on the use of programmable virtual scenarios to generate visual datasets and the needed associated annotations. For example, in an image-based machine learning technique, using a modern rendering engine (i.e., capable of producing photo-realistic imagery) has been proven a valid companion to automatically generate adequate datasets.

We successfully applied the *Virtual World approach* using the *Grand Tefth Auto V* engine for detection of personal protection equipment [7]), and for pedestrian detection (Sensors journal paper [17]). In particular, in [17] we considered the existing *Synthetic2Real* Domain Shift, and we tackled it exploiting two simple but effective Domain Adaptation approaches that try to create domain-invariant features.

1.5.2 Visual Counting

The counting problem is the estimation of the number of objects instances in still images or video frames. This task has recently become a hot research topic due to its interdisciplinary and widespread applicability and to its paramount importance for many real-world applications, for instance, counting bacterial cells from microscopic images, estimate the number of people present at an event, counting animals in ecological surveys with the intention of monitoring the population of a certain region, counting the number of trees in an aerial image of a forest, evaluate the number of vehicles in a highway or in a car park, monitoring crowds in surveillance systems, and others.

In humans, studies have demonstrated that, as a consequence of the subitizing ability , the brain switches between two techniques in order to count objects . When the observed objects are less than five, the fast and accurate Parallel Individuation System(PIS) is employed, otherwise, the inaccurate and error-prone Approximate Number System(ANS) is used. Thus, at least for crowded scenes, Computer Vision approaches offer a fast and useful alternative for counting objects.

In principle, the key idea behind objects counting using Computer Vision-based techniques is very simple: density times area. However, objects are not regular across the scene.They cluster in certain regions and are spread out in others. Another factor of complexity is represented by perspective distortions created by different camera viewpoints in various scenes, resulting in large variability of scales of objects. Others challenges points to be considered are inter-object and intra-object occlusions, high similarity of appearance between objects and background elements, different illuminations, and low image quality.

In order to overcome these challenges, several machine learning-based approaches (especially supervised and based on Convolutional Neural Networks) have been suggested. However, most of these methods require a large amount of labeled data and make a common assumption: the training and testing data are drawn from the same distribution. The direct transfer of the learned features between different domains does not work very well because the distributions are different. Thus, a model trained on one domain, named source, usually experiences a drastic drop in performance when applied on another domain, named target. This problem is commonly referred as *Domain Shift*.

Domain Adaptation is a common technique to address this problem. It adapts a trained neural network by fine-tuning it with a new set of labeled data belonging to the new distribution. In this way, we proposed some solutions able to count vehicles located in parking lots, fine-tuning and specializing some CNNs to work in this specific scenario. In particular, we introduced some detection-based approaches able to localize and count vehicles from images directly onboard smart-cameras and drones.

However, in many real cases, gathering a further collection of labeled data is expensive, especially for tasks that imply per-pixel annotations. Recently, we propose an end-to-end CNN-based *Unsupervised* Domain Adaptation algorithm for traffic density estimation and counting [18] that can generalize to new sources of data for which there is no training data available. We achieve this generalization by adversarial learning, whereby a discriminator attached to the output induces similar density distribution in the target and source domains.

1.5.3 Face Recognition

Face recognition is a key task in many application fields, such as security and surveillance. Several approaches have been proposed in the last few years to implement the face recognition task. Some approaches are based on local features of the images, such as Local Binary Pattern (LBP) which combines local descriptors of the face in order to build a global representation that can be used to measure the distance with other LBP features. Some other approaches are based on detecting the facial landmarks from the detected face and on measuring the distance between some of these landmarks. Recently, Deep Learning approach and Convolutional Neural Networks (CNNs) have been proposed to address the face verification problem with very good results.

We implemented several solutions based on the aforementioned techniques to address the face recognition problem in different application scenarios. For example, we studied the problem of intrusion detection in a monitored environment with embedded devices.

Recently, we started to perform experiments of face recognition on drone-acquired images [**?**, 4]. This is an even more challenging scenario, since the drones move, they are affected by weather conditions, and they are usually far from the monitored target, thus the resulting acquired images are often low-resolution, blurred and the face to be recognized is very small.

Although, deep models have shown impressive performance on the face recognition tasks, namely face verification and identification, their ability to recognize faces drastically reduce when dealing with low- and cross-resolution images. Concerning such an issue, we developed an algorithm to train models in a cross-resolution domain [31, 34]. With our algorithm we improved upon than the state-of-the-art, concerning the face recognition tasks on low- and cross-resolution domains, up to two orders of magnitude for images characterized by a resolutions from 32 down to 8 pixels (considering the shortest side).

1.6 Multimedia Information Retrieval

1.6.1 Video Browsing

Video data is the fastest growing data type on the Internet, and because of the proliferation of high-definition video cameras, the volume of video data is exploding. This data explosion in the video area has led to push research on large-scale video retrieval systems that are effective, fast, and easy to use for content search scenarios.

Within this framework, we developed a content-based video retrieval system VISIONE[1], to compete at the Video Browser Showdown (VBS), an international video search competition that evaluates the performance of interactive video retrievals systems. The tasks evaluated during the competition are: *Known-Item-Search (KIS)*, *textual KIS* and *Ad-hoc Video Search (AVS)*. The visual KIS task models the situation in which someone wants to find a particular video clip that he has already seen, assuming that it is contained in a specific collection of data. In the textual KIS, the target video clip is no longer visually presented to the participants of the challenge but it is rather described in details by text. This task simulates situations in which a user wants to find a particular video clip, without having seen it before, but knowing the content of the video exactly. For the AVS task, instead, a textual description is provided (e.g. "A person playing guitar outdoors") and participants need to find as many correct examples as possible, i.e. video shots that fit the given description.

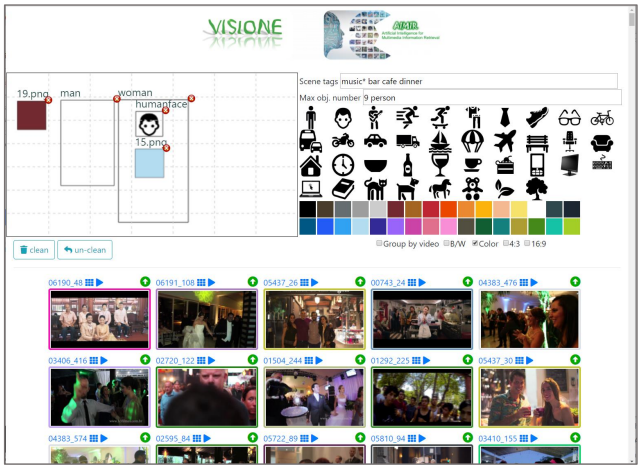

Figure 1. VISIONE User Interface

VISIONE can be used to solve both Known-Item and Ad-hoc Video Search tasks as it integrates several content-based analysis and retrieval modules, including a keyword search, a spatial object-based search, a spatial color-based search, and a visual similarity search. The user interface, shown in Figure 1, provides a text box to specify the keywords, and a canvas for sketching objects and colors to be found in the target video.

VISIONE is based on state-of-the-art deep learning approaches for the visual content analysis and exploits highly efficient indexing techniques to ensure scalability. In particular, it uses specifically designed textual encodings for indexing and searching video content. This aspect of our system is crucial: we can exploit the latest text search engine technologies, which nowadays are characterized by high efficiency and scalability, without the need to define a dedicated data structure or even worry about implementation issues.

A detailed description of all the functionalities included in VISIONE and how each of them are implemented is provided in [2]. Moreover, in [2] we presented an analysis of the system retrieval performance, by examining the logs acquired during the VBS 2019 challenge.

1.6.2 Similarity Search

Searching a data set for the most similar objects to a given query is a fundamental task in many branches of computer science, including pattern recognition, computational biology, and multimedia information retrieval, to name but a few. This search paradigm, referred to as *similarity search*, overcomes

[1] http://visione.isti.cnr.it/

limitations of traditional *exact-match search* that is neither feasible nor meaningful for complex data (e.g., multimedia data, vectorial data, time-series, etc.). In our research, we mainly focus on *metric search* methods, which are based on the assumption that data objects are represented as elements of a space (D,d) where the metric function d provides a measure of the closeness (i.e. dissimilarity) of the data objects. A proximity query is defined by a query object $q \in D$ and a proximity condition, such as "find all the objects within a threshold distance of q" (*range query*) or "finding the k closest objects to q" (*k-nearest neighbour query*). The exact response to a query is the set of all the data objects that satisfy the considered proximity condition.

Providing an exact response to a proximity query is not feasible if the search space is very large or it has a high intrinsic dimensionality since in such cases, the exact search rarely outperforms a sequential scan (phenomenon known as the *curse of dimensionality*). To overcome this issue, the research community has developed a wide spectrum of techniques for *approximate search*, which have higher efficiency though at the price of some imprecision in the results (e.g. some relevant results might be missing or some ranking errors might occur).

In the past, we developed and proposed various techniques to support approximate similarity research in metric spaces. Many of those techniques exploits the idea of transforming the original data objects into a more tractable space in which we can efficiently perform the search. For example, we proposed several *Permutation-Based Indexing* approaches where data objects are represented as a sequence of identifiers (*permutation*) that can be efficiently indexed and searched (e.g., by using inverted files), and *Sketching techniques*, which transform the data objects into compact binary strings. In the last years, we also investigated the use of some geometrical properties (namely, the 4-point property and the n-point property) to support metric search. For the class of metric space that satisfy the 4-point property, called *Supermetric spaces*, we derived a new pruning rule named *Hilbert Exclusion*, which can be used with any indexing mechanism based on hyperplane partitioning in order to determine subset of data that do not need to be exhaustively inspected. Moreover, for the large class of metric spaces meeting the n-point property (notably including Cartesian spaces of any dimension with the Euclidean, Cosine or Quadratic Form distances) we defined the *nSimplex projection* that allows mapping metric objects into a finite-dimensional Euclidean space where upper- and lower- bounds of the actual distance can be computed.

During 2020, we further investigated the use of the n-point property and the nSimplex projection for Approximate Nearest Neighbor search. In particular, in [52] we presented an approach that exploits a pivot-based local embedding to refine a set of candidate results of a similarity query. We focused our attention on refining of a set of approximate nearest neighbour results retrieved using a permutation-based search system. However, our approach can be generalized to other types of approximate search provided that they are based on the use of anchor objects (pivots) from which we precalculate the distances for other purposes. The core idea of the proposed technique is using the distances between an object and a set of pivots (pre-computed at indexing time) to embed the data objects into a low-dimensional where it is possible to compute upper- and lower-bounds for the actual distance (e.g. using the nSimplex projection). Dissimilarity functions defined upon those bounds are then adopted for re-ranking the candidate objects. The main advantage is that the proposed refining approach does not need to access the original data as done, instead, by the most commonly used refining technique that relies on computing the actual distances between the query and each candidate object.

1.6.3 Relational Cross-Modal Visual-Textual Retrieval

In the growing area of computer vision, modern deep-learning architectures are quite good at tasks such as classifying or recognizing objects in images. Recent studies, however, demonstrated the difficulties of such architectures to intrinsically understand a complex scene to catch spatial, temporal and abstract relationships among objects. Motivated by these limitations of the content-based information retrieval methods, we initially tackled the problem introducing a novel task, called R-CBIR (Relational Content-Based Image Retrieval). Given a query image, the objective of is catching images that are similar to the input query not only in terms of detected entities but also with respect to the relationships (spatial and non-spatial) between them. We experimented with different variations of the recently introduced Relation Network architecture to extract relationship-aware visual features. In particular, we approached the problem transferring knowledge from the Relation Network module trained on the R-VQA task using the CLEVR dataset. Under this setup, we initially introduced the Two-Stage Relation Network (2S-RN) and the Aggregated Visual Features Relation Network (AVF-RN) modules. The first introduces late-fusion of question features into the visual pipeline in order to produce visual features not conditioned on the particular question. In the latter, we solved the problem of producing a compact and representative visual relationship-aware feature by aggregating all the possible couples of objects directly inside the network for training it end to end.

During 2020, we extended this work to account for more realistic use-cases, concentrating the attention on real-world pictures. Furthermore, we addressed the problem of cross-modal visual-textual retrieval, which consists in finding pictures given a natural language description as a query (sentence-to-image retrieval) or vice-versa (image-to-sentence retrieval). In real-world search engines, these are very interesting and challenging scenarios. However, we initially tackled the sentence-to-image retrieval scenario, as it is the more attractive in real world use-cases. More in details, we introduced the Transformer Encoder Reasoning Network (TERN) [39], a deep relational neural network which is able to match images and sentences in a highly-semantic common space. The core

of the architecture is constituted of recently introduced deep relational modules called *transformer encoders*, which can spot out hidden intra-object relationships. We showed that this simple pipeline is able to create compact relational cross-modal descriptions that can be used for efficient similarity search.

More recently, we proposed an extension to TERN, called TERAN (Transformer Encoder Reasoning and Alignment Network) [38] which is able to obtain a fine-grained region-word alignment keeping the context into consideration. However, the network is still supervised at a global image-sentence level, and the fine-grained correspondences are automatically discovered. With this constraint during the learning phase, we obtained state-of-the-art results on the Recall@K metrics and on the novel NDCG metric with ROUGE-L and SPICE textual similarities used as relevances. This novel network effectively produces visually pleasant precise region-word alignments, and we also demonstrated how the fine-grained region-word alignment objective improves the retrieval effectiveness of the original TERN cross-modal descriptions.

The main motivations and preliminary results from these works are also available in the short paper [37] presented at the SISAP 2020 Doctoral Symposium. Most of the code for replicating the experiments is also available on GitHub (see Section 7.0.2 for more details).

2. Projects & Activities

2.1 EU Projects

AI4EU

In January 2019, the AI4EU consortium was established to build the first European Artificial Intelligence On-Demand Platform and Ecosystem with the support of the European Commission under the H2020 programme. The activities of the AI4EU project include:

- The creation and support of a large European ecosystem spanning the 28 countries to facilitate collaboration between all Europeans actors in AI (scientists, entrepreneurs, SMEs, Industries, funding organizations, citizens...);
- The design of a European AI on-Demand Platform to support this ecosystem and share AI resources produced in European projects, including high-level services, expertise in AI research and innovation, AI components and datasets, high-powered computing resources and access to seed funding for innovative projects using the platform;
- The implementation of industry-led pilots through the AI4EU platform, which demonstrates the capabilities of the platform to enable real applications and foster innovation; Research activities in five key interconnected AI scientific areas (Explainable AI, Physical AI ,Verifiable AI, Collaborative AI, Integrative AI), which arise from the application of AI in real-world scenarios;
- The funding of SMEs and start-ups benefitting from AI resources available on the platform (cascade funding plan of €3M) to solve AI challenges and promote new solutions with AI; The creation of a European Ethical Observatory to ensure that European AI projects adhere to high ethical, legal, and socio-economical standards;
- The production of a comprehensive Strategic Research Innovation Agenda for Europe; The establishment of an AI4EU Foundation that will ensure a handover of the platform in a sustainable structure that supports the European AI community in the long run.

The leader of the AIMH team participating in AI4EU is ...

AI4Media

Artificial Intelligence for the Society and the Media Industry (AI4Media) is a network of research excellence centres delivering advances in AI technology in the media sector. Funded under H2020-EU.2.1.1., AI4Media started in September 2020 and will end in August 2024.

Motivated by the challenges, risks and opportunities that the wide use of AI brings to media, society and politics, AI4Media aspires to become a centre of excellence and a wide network of researchers across Europe and beyond, with a focus on delivering the next generation of core AI advances to serve the key sector of Media, to make sure that the European values of ethical and trustworthy AI are embedded in future AI deployments, and to reimagine AI as a crucial beneficial enabling technology in the service of Society and Media.

The leader of the AIMH team participating in AI4Media is Fabrizio Sebastiani.

ARIADNEplus

The ARIADNEplus project is the extension of the previous ARIADNE Integrating Activity, which successfully integrated archaeological data infrastructures in Europe, indexing in its registry about 2.000.000 datasets. ARIADNEplus will build on the ARIADNE results, extending and supporting the research community that the previous project created and further developing the relationships with key stakeholders such as the most important European archaeological associations, researchers, heritage professionals, national heritage agencies and so on. The new enlarged partnership of ARIADNEplus covers all of Europe. It now includes leaders in different archaeological domains like palaeoanthropology, bioarchaeology and environmental archaeology as well as other sectors of archaeological sciences, including all periods of human presence from the appearance of hominids to present times.

Transnational Activities together with the planned training will further reinforce the presence of ARIADNEplus as a key actor. The technology underlying the project is state-of-art. The ARIADNEplus data infrastructure will be embedded in a cloud that will offer the availability of Virtual Research Environments where data-based archaeological research may be carried out. The project will furthermore develop a Linked Data approach to data discovery. Innovative services will be made available to users, such as visualization, annotation, text mining and geo-temporal data management. Innovative pilots will be developed to test and demonstrate the innovation potential of the ARIADNEplus approach. Fostering innovation will be a key aspect of the project, with dedicated activities led by the project Innovation Manager.

Mingei

The Mingei Project explores the possibilities of representing and making accessible both tangible and intangible aspects of craft as cultural heritage (CH). Heritage Crafts (HCs) involve craft artefacts, materials, and tools and encompass craftsmanship as a form of Intangible Cultural Heritage. Intangible HC dimensions include dexterity, know-how, and skilled use of tools, as well as, tradition, and identity of the communities in which they are, or were, practiced. HCs are part of the history and have impact upon the economy of the areas in which they flourish. The significance and urgency to the preservation of HCs is underscored, as several are threatened with extinction. Despite their cultural significance efforts for HC representation and preservation are scattered geographically and thematically. Mingei provides means to establish HC representations based on digital assets, semantics, existing literature and repositories, as well as, mature digitisation and representation technologies. These representations will capture and preserve tangible and intangible dimensions of HCs. Central to craftsmanship is skill and its transmission from master to apprentice. Mingei captures the motion and tool usage of HC practitioners, from Living Human Treasures and archive documentaries, in order to preserve and illustrate skill and tool manipulation. The represented knowledge will be availed through experiential presentations, using storytelling and educational applications and based on Advanced Reality, Mixed Reality and the Internet. The project has started on December 1, 2019 and will last 3 years.

MultiForesee

The main objective of this Action, entitled MULTI-modal Imaging of FORensic SciEnce Evidence (MULTI-FORESEE)-tools for Forensic Science[2], is to promote innovative, multi-informative, operationally deployable and commercially exploitable imaging solutions/technology to analyse forensic evidence.

Forensic evidence includes, but not limited to, fingermarks, hair, paint, biofluids, digital evidence, fibers, documents and living individuals. Imaging technologies include optical, mass spectrometric, spectroscopic, chemical, physical and digital forensic techniques complemented by expertise in IT solutions and computational modelling.

Imaging technologies enable multiple physical and chemical information to be captured in one analysis, from one specimen, with information being more easily conveyed and understood for a more rapid exploitation. The enhanced value of the evidence gathered will be conducive to much more informed investigations and judicial decisions thus contributing to both savings to the public purse and to a speedier and stronger criminal justice system.

The Action will use the unique networking and capacity-building capabilities provided by the COST framework to bring together the knowledge and expertise of Academia, Industry and End Users. This synergy is paramount to boost imaging technological developments which are operationally deployable.

The leader of the AIMH team participating in MultiForesee is Giuseppe Amato.

SoBigData++

SoBigData++ is a project funded by the European Commission under the H2020 Programme INFRAIA-2019-1, started Jan 1 2020 and ending Dec 31, 2023. SoBigData++ proposes to create the Social Mining and Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data". SoBigData plans to open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research. It plans to not only strengthen the existing clusters of excellence in social data mining research, but also create a pan-European, inter-disciplinary community of social data scientists, fostered by extensive training, networking, and innovation activities.

The leader of the AIMH team participating in SoBigData++ is Fabrizio Sebastiani.

2.2 SSHOC

Social Sciences Humanities Open Cloud (SSHOC) is a project funded by the EU framework programme Horizon 2020 and unites 20 partner organisations and their 27 associates in developing the social sciences and humanities area of the European Open Science Cloud (EOSC). SSHOC partners include both

[2] https://multiforesee.com/

developing and fully established European Research Infrastructures from the social sciences and humanities, and the association of European research libraries (LIBER). The goal of the project is to transform the social sciences humanities data landscape with its disciplinary silos and separate facilities into an integrated, cloud-based network of interconnected data infrastructures. To promote synergies and open science initiatives between disciplines, and accelerate interdisciplinary research and collaboration, these data infrastructures will be supported by the tools and training which allow scholars and researchers to access, process, analyse, enrich and compare data across the boundaries of individual repositories or institutions. SSHOC will continuously monitor ongoing developments in the EOSC so as to conform to the necessary technical and other requirements for making the SSHOC services sustainable beyond the duration of the project. Some of the results obtained by the AIMH team involved in SSHOC have been presented in [NN] The leader of the AIMH team participating in SSHOC is Cesare Concordia. https://sshopencloud.eu

2.3 CNR National Virtual Lab on AI

Fabrizio Falchi has coordinated, together with Sara Colantonio, the activities of the National Virtual Lab of CNR on Artificial Intelligence. This initiative connects about 90 groups in 22 research institutes of 6 departments of the whole CNR. The Nationtal Virtual Lab on AI aims at proposing a strategic vision and big and long-term projects.

2.4 National Projects

AI-MAP

AI-MAP is a project funded by Regione Toscana that aims at analyzing digitized historical geographical regional maps using deep learning methods to increase the availability and searchability of the digitized documents. The main objectives of the project is to develop automatic or semi-automatic pipelines for denoising/repairing of the digitized documents, handwritten toponym localization and transcription. The activities are mainly conducted in the context of this project by Fabio Carrara under the scientific coordination of Giuseppe Amato.

AI4CHSites

AI4CHSites is a project funded by Regione Toscana that aims at analyzing visual content from surveillance camera in a touristic scenario. Partners of the project are: Opera della Primaziale Pisana and INERA srl. The activities in the context of this project are mainly conducted by Nicola Messina under the scientific coordination of Fabrizio Falchi.

ADA

In the era of Big Data, manufacturing companies are overwhelmed by a lot of disorganized information: the large amount of digital content that is increasingly available in the manufacturing process makes the retrieval of accurate information a critical issue. In this context, and thanks also to the Industry 4.0 campaign, the Italian manufacturing industries have made a lot of effort to ameliorate their knowledge management system using the most recent technologies, like big data analysis and machine learning methods. In this context, therefore, the main target of the ADA project is to design and develop a platform based on big data analytics systems that allows for the acquisition, organization, and automatic retrieval of information from technical texts and images in the different phases of acquisition, design & development, testing, installation and maintenance of products.

HDN

Hypermedia Dante Network (HDN) is a three year (2020-2023) Italian National Research Project (PRIN) which aims to extend the ontology and tools developed by AIMH team to represent the sources of Dante Alighieri's minor works to the more complex world of the Divine Comedy. In particular, HDN aims to enrich the functionalities of the DanteSources Web application (https://dantesources.dantenetwork.it/) in order to efficiently recover knowledge about the Divine Comedy. Relying on some of the most important scientific institutions for Dante studies, such as the Italian Dante Society of Florence, HDN makes use of specialized skills, essential for the population of ontology and the consequent creation of a complete and reliable knowledge base. Knowledge will be published on the Web as Linked Open Data and will be access through a user-friendly Web application.

IMAGO

The IMAGO (Index Medii Aevi Geographiae Operum) is a three year (2020-2023) Italian National Research Project (PRIN) that aims at creating a knowledge base of the critical editions of Medieval and Humanistic Latin geographical works (VI-XV centuries). Up to now, this knowledge has been collected in many paper books or several databases, making it difficult for scholars to retrieve it easily and to produce a complete overview of these data. The goal of the project is to develop new tools that satisfy the needs of the academic research community, especially for scholars interested in Medieval and Renaissance Humanism geography. Using Semantic Web technologies, AIMH team will develop an ontology providing the terms to represent this knowledge in a machine-readable form. A semi-automatic tool will help the scholars to populate the ontology with the data included in authoritative critical editions. Afterwards, the tool will automatically save the resulting graph into a triple store. On top of this graph, a Web application will be developed, which will allow users to extract and display the information stored in the knowledge base in the form of maps, charts, and tables.

VIDEMO

Visual Deep Engines for Monitoring (VIDEMO) is a 2-year project funded by Regione Toscana, Istituto di Scienza e Tecnologie dell'Informazione "A.Faedo" (ISTI) del CNR, Visual Engines srl. VIDEMO is about automatic analysis of images and video using deep learning methods for secure societies.

The activities reported in Section 1.5.3 have been mainly conducted in the context of this project by Fabio Valerio Massoli. Fabrizio Falchi is the scientific coordinator of the project.

*WAC@Lucca WeAreClouds @Lucca carries out research and development activities in the field of monitoring public places, such as squares and streets, through cameras and microphones with artificial intelligence technologies, in order to collect useful information both for the evaluation of tourist flows and their impact. on the city, both for purposes of automatic identification of particular events of interest for statistical purposes or for security. The project is funded by Fondazione Cassa di Risparmio di Lucca and Comune di Lucca is a partner. Fabrizio Falchi is the scientific coordinator of the project.

3. Papers

In this section, we report the complete list of paper we published in 2020 organized in four categories: journals, proceedings, magazines, others, and preprints.

3.1 Journals

In this section, we report the paper we published (or accepted for publication) in journals during 2020, in alphabetic order of the first author.

3.1.1
Large-scale instance-level image retrieval

G. Amato, F. Carrara, F. Falchi, C. Gennaro, L. Vadicamo.

In Elsevier, Information Processing & Management special issue on Deep Learning for Information Retrieval. [3].

The great success of visual features learned from deep neural networks has led to a significant effort to develop efficient and scalable technologies for image retrieval. Nevertheless, its usage in large-scale Web applications of content-based retrieval is still challenged by their high dimensionality. To overcome this issue, some image retrieval systems employ the product quantization method to learn a large-scale visual dictionary from a training set of global neural network features. These approaches are implemented in main memory, preventing their usage in big-data applications. The contribution of the work is mainly devoted to investigating some approaches to transform neural network features into text forms suitable for being indexed by a standard full-text retrieval engine such as Elasticsearch. The basic idea of our approaches relies on a transformation of neural network features with the twofold aim of promoting the sparsity without the need of unsupervised pre-training. We validate our approach on a recent convolutional neural network feature, namely Regional Maximum Activations of Convolutions (R-MAC), which is a state-of-art descriptor for image retrieval. Its effectiveness has been proved through several instance-level retrieval benchmarks. An extensive experimental evaluation conducted on the standard benchmarks shows the effectiveness and efficiency of the proposed approach and how it compares to state-of-the-art main-memory indexes.

3.1.2
Efficient Evaluation of Image Quality via Deep-Learning Approximation of Perceptual Metrics

A. Artusi, F. Banterle, F. Carrara, A. Moreo.

In IEEE Transactions on Image Processing, vol. 29. [5].

Image metrics based on Human Visual System (HVS) play a remarkable role in the evaluation of complex image processing algorithms. However, mimicking the HVS is known to be complex and computationally expensive (both in terms of time and memory), and its usage is thus limited to a few applications and to small input data. All of this makes such metrics not fully attractive in real-world scenarios. To address these issues, we propose Deep Image Quality Metric *(DIQM), a deep-learning approach to learn the global image quality feature (*mean-opinion-score*). DIQM can emulate existing visual metrics efficiently, reducing the computational costs by more than an order of magnitude with respect to existing implementations.*

3.1.3
Learning accurate personal protective equipment detection from virtual worlds

M. Di Benedetto and F. Carrara and E. Meloni and G. Amato and F. Falchi and C. Gennaro

In Springer, Multimedia Tools and Applications. [7].

Deep learning has achieved impressive results in many machine learning tasks such as image recognition and computer vision. Its applicability to supervised problems is however constrained by the availability of high-quality training data consisting of large numbers of humans annotated examples (e.g. millions). To overcome this problem, recently, the AI world is increasingly exploiting artificially generated images or video sequences using realistic photo rendering engines such as those used in entertainment applications. In this way, large sets of training images can be easily created to train deep learning algorithms. In this paper, we generated photo-realistic synthetic image sets to train deep learning models to recognize the correct use of personal safety equipment (e.g., worker safety helmets, high visibility vests, ear protection devices) during at-risk work activities. Then, we performed the adaptation of the domain to real-world images using a very small set of real-world images. We demonstrated that training with the synthetic training set generated and the use of the domain adaptation phase is an effective solution for applications where no training set is available.

3.1.4
Virtual to Real Adaptation of Pedestrian Detectors

Luca Ciampi, Nicola Messina, Fabrizio Falchi, Claudio Gennaro and Giuseppe Amato

In NDPI, Sensors vol. 20(18). [17].

Pedestrian detection through Computer Vision is a building block for a multitude of applications. Recently, there has been an increasing interest in convolutional neural network-based architectures to execute such a task. One of these supervised networks' critical goals is to generalize the knowledge learned during the training phase to new scenarios with different characteristics. A suitably labeled dataset is essential to achieve this purpose. The main problem is that manually annotating a dataset usually requires a lot of human effort, and it is costly. To this end, we introduce ViPeD (Virtual Pedestrian Dataset), a new synthetically generated set of images collected

with the highly photo-realistic graphical engine of the video game GTA V (Grand Theft Auto V), where annotations are automatically acquired. However, when training solely on the synthetic dataset, the model experiences a Synthetic2Real domain shift leading to a performance drop when applied to real-world images. To mitigate this gap, we propose two different domain adaptation techniques suitable for the pedestrian detection task, but possibly applicable to general object detection. Experiments show that the network trained with ViPeD can generalize over unseen real-world scenarios better than the detector trained over real-world data, exploiting the variety of our synthetic dataset. Furthermore, we demonstrate that with our domain adaptation techniques, we can reduce the Synthetic2Real domain shift, making the two domains closer and obtaining a performance improvement when testing the network over the real-world images.

3.1.5
A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment

A. Esuli, A. Molinari, F. Sebastiani.

In ACM Transactions on Information Systems vol. 39(2), 2020. [21].

We critically re-examine the Saerens-Latinne-Decaestecker (SLD) algorithm, a well-known method for estimating class prior probabilities ("priors") and adjusting posterior probabilities ("posteriors") in scenarios characterized by distribution shift, i.e., difference in the distribution of the priors between the training and the unlabelled documents. Given a machine-learned classifier and a set of unlabelled documents for which the classifier has returned posterior probabilities and estimates of the prior probabilities, SLD updates them both in an iterative, mutually recursive way, with the goal of making both more accurate; this is of key importance in downstream tasks such as single-label multiclass classification and cost-sensitive text classification. Since its publication, SLD has become the standard algorithm for improving the quality of the posteriors in the presence of distribution shift, and is still considered a top contender when we need to estimate the priors (a task that has become known as "quantification"). However, its real effectiveness in improving the quality of the posteriors has been questioned. We here present the results of systematic experiments conducted on a large, publicly available dataset, across multiple amounts of distribution shift and multiple learners. Our experiments show that SLD improves the quality of the posterior probabilities and of the estimates of the prior probabilities, but only when the number of classes in the classification scheme is very small and the classifier is calibrated. As the number of classes grows, or as we use non-calibrated classifiers, SLD converges more slowly (and often does not converge at all), performance degrades rapidly, and the impact of SLD on the quality of the prior estimates and of the posteriors becomes negative rather than positive.

3.1.6
Cross-lingual sentiment quantification

A. Esuli, A. Moreo, F. Sebastiani.

In IEEE Intelligent Systems, vol. 35, 2020. [22].

Sentiment Quantification *(i.e., the task of estimating the relative frequency of sentiment-related classes — such as* Positive *and* Negative *— in a set of unlabelled documents) is an important topic in sentiment analysis, as the study of sentiment-related quantities and trends across a population is often of higher interest than the analysis of individual instances. In this work we propose a method for* Cross-Lingual Sentiment Quantification*, the task of performing sentiment quantification when training documents are available for a source language $\mathscr{S}$ but not for the target language $\mathscr{T}$ for which sentiment quantification needs to be performed. Cross-lingual sentiment quantification (and cross-lingual* text *quantification in general) has never been discussed before in the literature; we establish baseline results for the binary case by combining state-of-the-art quantification methods with methods capable of generating cross-lingual vectorial representations of the source and target documents involved. We present experimental results obtained on publicly available datasets for cross-lingual sentiment classification; the results show that the presented methods can perform cross-lingual sentiment quantification with a surprising level of accuracy.*

3.1.7
5G-Enabled Security Scenarios for Unmanned Aircraft: Experimentation in Urban Environment

E. Ferro, C. Gennaro, A. Nordio, F. Paonessa, C. Vairo, G. Virone, A. Argentieri, A. Berton, A. Bragagnini

In MDPI, Drones, 2020, 4.2: 22. [**?**].

The telecommunication industry has seen rapid growth in the last few decades. This trend has been fostered by the diffusion of wireless communication technologies. In the city of Matera, Italy (European capital of culture 2019), two applications of 5G for public security have been tested by using an aerial drone: the recognition of objects and people in a crowded city and the detection of radio-frequency jammers. This article describes the experiments and the results obtained. The drone flew at a height of 40m, never on people and in weather conditions with strong winds. The results obtained on facial recognition are therefore exceptional, given the conditions in which the data were found.

3.1.8
Cross-resolution learning for Face Recognition

F.V. Massoli, G. Amato, F. Falchi

In Elsevier, Image and Vision Computing, vol. 99. [31].

Deep learning, Low resolution Face Recognition, Cross resolution Face Recognition", abstract = "Convolutional Neural Network models have reached extremely high performance on the Face Recognition task. Mostly used datasets, such as VGGFace2, focus on gender, pose, and age variations, in the attempt of balancing them to empower models to better generalize to unseen data. Nevertheless, image resolution variability is not usually discussed, which may lead to a resizing of 256 pixels. While specific datasets for very low-resolution faces have been proposed, less attention has been paid on the task of cross-resolution matching. Hence, the discrimination power of a neural network might seriously degrade in such a scenario. Surveillance systems and forensic applications are particularly susceptible to this problem since, in these cases, it is common that a low-resolution query has to be matched against higher-resolution galleries. Although it is always possible to either

increase the resolution of the query image or to reduce the size of the gallery (less frequently), to the best of our knowledge, extensive experimentation of cross-resolution matching was missing in the recent deep learning-based literature. In the context of low- and cross-resolution Face Recognition, the contribution of our work is fourfold: i) we proposed a training procedure to fine-tune a state-of-the-art model to empower it to extract resolution-robust deep features; ii) we conducted an extensive test campaign by using high-resolution datasets (IJB-B and IJB-C) and surveillance-camera-quality datasets (QMUL-SurvFace, TinyFace, and SCface) showing the effectiveness of our algorithm to train a resolution-robust model; iii) even though our main focus was the cross-resolution Face Recognition, by using our training algorithm we also improved upon state-of-the-art model performances considering low-resolution matches; iv) we showed that our approach could be more effective concerning preprocessing faces with super-resolution techniques. The python code of the proposed method will be available at https://github.com/fvmassoli/cross-resolution-face-recognition.

3.1.9
Detection of Face Recognition Adversarial Attacks

F.V. Massoli, F. Carrara, G. Amato, F. Falchi

Elsevier, Computer Vision and Image Understanding Volume 202, 103103. [32]

Deep Learning methods have become state-of-the-art for solving tasks such as Face Recognition (FR). Unfortunately, despite their success, it has been pointed out that these learning models are exposed to adversarial inputs – images to which an imperceptible amount of noise for humans is added to maliciously fool a neural network – thus limiting their adoption in sensitive real-world applications. While it is true that an enormous effort has been spent to train robust models against this type of threat, adversarial detection techniques have recently started to draw attention within the scientific community. The advantage of using a detection approach is that it does not require to re-train any model; thus, it can be added to any system. In this context, we present our work on adversarial detection in forensics mainly focused on detecting attacks against FR systems in which the learning model is typically used only as features extractor. Thus, training a more robust classifier might not be enough to counteract the adversarial threats. In this frame, the contribution of our work is four-fold: (i) we test our proposed adversarial detection approach against classification attacks, i.e., adversarial samples crafted to fool an FR neural network acting as a classifier; (ii) using a k-Nearest Neighbor (k-NN) algorithm as a guide, we generate deep features attacks against an FR system based on a neural network acting as features extractor, followed by a similarity-based procedure which returns the query identity; (iii) we use the deep features attacks to fool an FR system on the 1:1 face verification task, and we show their superior effectiveness with respect to classification attacks in evading such type of system; (iv) we use the detectors trained on the classification attacks to detect the deep features attacks, thus showing that such approach is generalizable to different classes of offensives.

3.1.10
Cross-resolution face recognition adversarial attacks

F.V. Massoli and F. Falchi and G. Amato

In Elsevier, Pattern Recognition Letters, vol. 140, pp. 222-229. [33]

Face Recognition is among the best examples of computer vision problems where the supremacy of deep learning techniques compared to standard ones is undeniable. Unfortunately, it has been shown that they are vulnerable to adversarial examples - input images to which a human imperceptible perturbation is added to lead a learning model to output a wrong prediction. Moreover, in applications such as biometric systems and forensics, cross-resolution scenarios are easily met with a non-negligible impact on the recognition performance and adversary's success. Despite the existence of such vulnerabilities set a harsh limit to the spread of deep learning-based face recognition systems to real-world applications, a comprehensive analysis of their behavior when threatened in a cross-resolution setting is missing in the literature. In this context, we posit our study, where we harness several of the strongest adversarial attacks against deep learning-based face recognition systems considering the cross-resolution domain. To craft adversarial instances, we exploit attacks based on three different metrics, i.e., L_1, L_2, and L_∞, and we study the resilience of the models across resolutions. We then evaluate the performance of the systems against the face identification protocol, open- and close-set. In our study, we find that the deep representation attacks represents a much dangerous menace to a face recognition system than the ones based on the classification output independently from the used metric. Furthermore, we notice that the input image's resolution has a non-negligible impact on an adversary's success in deceiving a learning model. Finally, by comparing the performance of the threatened networks under analysis, we show how they can benefit from a cross-resolution training approach in terms of resilience to adversarial attacks.

3.1.11
Representing Narratives in Digital Libraries: The Narrative Ontology

C. Meghini, V. Bartalesi, D. Metilli.

In IOS, Semantic Web Journal, Special Issue Cultural Heritage 2019. [36] *Digital Libraries (DLs), especially in the Cultural Heritage domain, are rich in narratives. Every digital object in a DL tells some kind of story, regardless of the medium, the genre, or the type of the object. However, DLs do not offer services about narratives, for example it is not possible to discover a narrative, to create one, or to compare two narratives. Certainly, DLs offer discovery functionalities over their contents, but these services merely address the objects that carry the narratives (e.g. books, images, audiovisual objects), without regard for the narratives themselves. The present work aims at introducing narratives as first-class citizens in DLs, by providing a formal expression of what a narrative is. In particular, this paper presents a conceptualization of the domain of narratives, and its specification through the Narrative Ontology (NOnt for short), expressed in first-order logic. NOnt has been implemented as an extension of three standard vocabularies, i.e. the CIDOC CRM, FRBRoo, and OWL Time, and using the SWRL rule language to express the axioms. An initial validation of NOnt has been performed in*

the context of the Mingei European project, in which the ontology has been applied to the representation of knowledge about Craft Heritage.

3.1.12
Learning to weight for text classification

A. Moreo, A. Esuli, F. Sebastiani.

In IEEE Transactions on Knowledge and Data Engineering, vol. 32, 2020. [41].

In information retrieval (IR) and related tasks, term weighting approaches typically consider the frequency of the term in the document and in the collection in order to compute a score reflecting the importance of the term for the document. In tasks characterized by the presence of training data (such as text classification) it seems logical that the term weighting function should take into account the distribution (as estimated from training data) of the term across the classes of interest. Although "supervised term weighting" approaches that use this intuition have been described before, they have failed to show consistent improvements. In this article we analyse the possible reasons for this failure, and call consolidated assumptions into question. Following this criticism we propose a novel supervised term weighting approach that, instead of relying on any predefined formula, learns a term weighting function optimised on the training set of interest; we dub this approach Learning to Weight *(LTW). The experiments that we run on several well-known benchmarks, and using different learning methods, show that our method outperforms previous term weighting approaches in text classification.*

3.1.13
Word-class embeddings for multiclass text classification

A. Moreo, A. Esuli, F. Sebastiani.

Data Mining and Knowledge Discovery. Forthcoming. [42].

Pre-trained word embeddings encode general word semantics and lexical regularities of natural language, and have proven useful across many NLP tasks, including word sense disambiguation, machine translation, and sentiment analysis, to name a few. In supervised tasks such as multiclass text classification (the focus of this article) it seems appealing to enhance word representations with ad-hoc embeddings that encode task-specific information. We propose (supervised) word-class embeddings *(WCEs), and show that, when concatenated to (unsupervised) pre-trained word embeddings, they substantially facilitate the training of deep-learning models in multiclass classification by topic. We show empirical evidence that WCEs yield a consistent improvement in multiclass classification accuracy, using six popular neural architectures and six widely used and publicly available datasets for multiclass text classification. One further advantage of this method is that it is conceptually simple and straightforward to implement. Our code that implements WCEs is publicly available at* `https://github.com/AlexMoreo/word-class-embeddings`.

3.1.14
Evaluation measures for quantification: An axiomatic approach

F. Sebastiani.

In Springer, Information Retrieval Journal, vol. 23. [49].

Quantification is the task of estimating, given a set σ of unlabelled items and a set of classes $\mathscr{C} = \{c_1, \ldots, c_{|\mathscr{C}|}\}$, the prevalence (or "relative frequency") in σ of each class $c_i \in \mathscr{C}$. While quantification may in principle be solved by classifying each item in σ and counting how many such items have been labelled with c_i, it has long been shown that this "classify and count" (CC) method yields suboptimal quantification accuracy. As a result, quantification is no longer considered a mere byproduct of classification, and has evolved as a task of its own. While the scientific community has devoted a lot of attention to devising more accurate quantification methods, it has not devoted much to discussing what properties an evaluation measure for quantification *(EMQ) should enjoy, and which EMQs should be adopted as a result. This paper lays down a number of interesting properties that an EMQ may or may not enjoy, discusses if (and when) each of these properties is desirable, surveys the EMQs that have been used so far, and discusses whether they enjoy or not the above properties. As a result of this investigation, some of the EMQs that have been used in the literature turn out to be severely unfit, while others emerge as closer to what the quantification community actually needs. However, a significant result is that no existing EMQ satisfies all the properties identified as desirable, thus indicating that more research is needed in order to identify (or synthesize) a truly adequate EMQ.*

3.1.15
Re-ranking via local embeddings: A use case with permutation-based indexing and the nSimplex projection

L. Vadicamo, C. Gennaro, F. Falchi, E. Chávez, R. Connor, G. Amato. In Elsevier, Information Systems, vol. 95. [52].

In this approach, the entire database is ranked by a permutation distance to the query. Typically, permutations allow the efficient selection of a candidate set of results, but typically to achieve high recall or precision this set has to be reviewed using the original metric and data. This can lead to a sizeable percentage of the database being recalled, along with many expensive distance calculations. To reduce the number of metric computations and the number of database elements accessed, we propose here a re-ranking based on a local embedding using the nSimplex projection.

The nSimplex projection produces Euclidean vectors from objects in metric spaces which possess the n-point property. The mapping is obtained from the distances to a set of reference objects, and the original metric can be lower bounded and upper bounded by the Euclidean distance of objects sharing the same set of references. Our approach is particularly advantageous for extensive databases or expensive metric function. We reuse the distances computed in the permutations in the first stage, and hence the memory footprint of the index is not increased. An extensive experimental evaluation of our approach is presented, demonstrating excellent results even on a set of hundreds of millions of objects.*Approximate Nearest Neighbor (ANN) search is a prevalent paradigm for searching intrinsically high dimensional objects in large-scale data sets. Recently, the permutation-based approach for ANN has attracted a lot of interest due to its versatility in being used in the more general class of metric spaces.*

In this approach, the entire database is ranked by a permutation distance to the query. Typically, permutations allow the efficient

selection of a candidate set of results, but typically to achieve high recall or precision this set has to be reviewed using the original metric and data. This can lead to a sizeable percentage of the database being recalled, along with many expensive distance calculations. To reduce the number of metric computations and the number of database elements accessed, we propose here a re-ranking based on a local embedding using the nSimplex projection.

The nSimplex projection produces Euclidean vectors from objects in metric spaces which possess the n-point property. The mapping is obtained from the distances to a set of reference objects, and the original metric can be lower bounded and upper bounded by the Euclidean distance of objects sharing the same set of references. Our approach is particularly advantageous for extensive databases or expensive metric function. We reuse the distances computed in the permutations in the first stage, and hence the memory footprint of the index is not increased. An extensive experimental evaluation of our approach is presented, demonstrating excellent results even on a set of hundreds of millions of objects.

3.1.16
MARC: a robust method for multiple-aspect trajectory classification via space, time, and semantic embeddings

L.M. Petry, C.L. Da Silva, A. Esuli, C. Renso, V. Bogorny.
In International Journal of Geographical Information Science, 34:7. [47].

The increasing popularity of Location-Based Social Networks (LBSNs) and the semantic enrichment of mobility data in several contexts in the last years has led to the generation of large volumes of trajectory data. In contrast to GPS-based trajectories, LBSN and context-aware trajectories are more complex data, having several semantic textual dimensions besides space and time, which may reveal interesting mobility patterns. For instance, people may visit different places or perform different activities depending on the weather conditions. These new semantically rich data, known as multiple-aspect trajectories, pose new challenges in trajectory classification, which is the problem that we address in this paper. Existing methods for trajectory classification cannot deal with the complexity of heterogeneous data dimensions or the sequential aspect that characterizes movement. In this paper we propose MARC, an approach based on attribute embedding and Recurrent Neural Networks (RNNs) for classifying multiple-aspect trajectories, that tackles all trajectory properties: space, time, semantics, and sequence. We highlight that MARC exhibits good performance especially when trajectories are described by several textual/categorical attributes. Experiments performed over four publicly available datasets considering the Trajectory-User Linking (TUL) problem show that MARC outperformed all competitors, with respect to accuracy, precision, recall, and F1-score.

3.1.17
Representation and Preservation of Heritage Crafts

X. Zabulis, C. Meghini, N. Partarakis, C. Beisswenger, A. Dubois, M. Fasoula, V. Nitti, S. Ntoa, I. Adami, A. Chatziantoniou, V. Bartalesi, D. Metilli, N. Stivaktakis, N. Patsiouras, P. Doulgeraki, E. Karuzaki, E. Stefanidi, A. Qammaz, D. Kaplanidi, I. Neumann-Janßen, U. Denter, H. Hauser, A. Petraki, I. Stivaktakis, E. Mantinaki, A. Rigaki and G. Galanakis.
In MDPI Sustainability, 12(4), 1461, 2020.

This work regards the digital representation of tangible and intangible dimensions of heritage crafts, towards craft preservation. Based on state-of-the-art digital documentation, knowledge representation and narrative creation approach are presented. Craft presentation methods that use the represented content to provide accurate, intuitive, engaging, and educational ways for HC presentation and appreciation are proposed. The proposed methods aim to contribute to HC preservation, by adding value to the cultural visit, before, and after it. [53]

3.2 Proceedings

In this section, we report the paper we published in proceedings of conferences during 2020 in alphabetic order of the first author.

3.2.1
Scalar Quantization-Based Text Encoding for Large Scale Image Retrieval

G. Amato, F. Carrara, F. Falchi, C. Gennaro, F. Rabitti, L. Vadicamo.
In 28th Italian Symposium on Advanced Database Systems, SEBD 2020, CEUR Workshop Proceedings, 2020, 2646, pp. 258-265. [1]

The great success of visual features learned from deep neural networks has led to a significant effort to develop efficient and scalable technologies for image retrieval. This paper presents an approach to transform neural network features into text codes suitable for being indexed by a standard full-text retrieval engine such as Elasticsearch. The basic idea is providing a transformation of neural network features with the twofold aim of promoting the sparsity without the need of unsupervised pre-training. We validate our approach on a recent convolutional neural network feature, namely Regional Maximum Activations of Convolutions (R-MAC), which is a state-of-art descriptor for image retrieval. An extensive experimental evaluation conducted on standard benchmarks shows the effectiveness and efficiency of the proposed approach and how it compares to state-of-the-art main-memory indexes.

3.2.2
Multi-Resolution Face Recognition with Drones

G. Amato, F. Falchi, C. Gennaro, F.V. Massoli, C. Vairo.
In 3rd International Conference on Sensors, Signal and Image Processing, SSIP 2020. [4]

Smart cameras have recently seen a large diffusion and represent a low-cost solution for improving public security in many scenarios. Moreover, they are light enough to be lifted by a drone. Face recognition enabled by drones equipped with smart cameras has already been reported in the literature. However, the use of the drone generally imposes tighter constraints than other facial recognition scenarios. First, weather conditions, such as the presence of wind, pose a severe limit on image stability. Moreover, the distance the drones fly is typically much high than fixed ground cameras, which inevitably translates into a degraded resolution of the face images. Furthermore, the drones' operational altitudes usually require the use of optical zoom, thus amplifying the harmful effects

of their movements. For all these reasons, in drone scenarios, image degradation strongly affects the behavior of face detection and recognition systems. In this work, we studied the performance of deep neural networks for face re-identification specifically designed for low-quality images and applied them to a drone scenario using a publicly available dataset known as DroneSURF.

3.2.3
Nor-Vdpnet: A No-Reference High Dynamic Range Quality Metric Trained On Hdr-Vdp 2

Francesco Banterle, Alessandro Artusi, Alejandro Moreo, Fabio Carrara.
In the 27th IEEE International Conference on Image Processing (ICIP) 2020, pp. 126-130. [6].

HDR-VDP 2 has convincingly shown to be a reliable metric for image quality assessment, and it is currently playing a remarkable role in the evaluation of complex image processing algorithms. However, HDR-VDP 2 is known to be computationally expensive (both in terms of time and memory) and is constrained to the availability of a ground-truth image (the so-called reference) against to which the quality of a processed imaged is quantified. These aspects impose severe limitations on the applicability of HDR-VDP 2 to realworld scenarios involving large quantities of data or requiring real-time responses. To address these issues, we propose Deep No-Reference Quality Metric (NoR-VDPNet), a deeplearning approach that learns to predict the global image quality feature (i.e., the mean-opinion-score index Q) that HDRVDP 2 computes. NoR-VDPNet is no-reference (i.e., it operates without a ground truth reference) and its computational cost is substantially lower when compared to HDR-VDP 2 (by more than an order of magnitude). We demonstrate the performance of NoR-VDPNet in a variety of scenarios, including the optimization of parameters of a denoiser and JPEG-XT.

3.2.4
Continuous ODE-Defined Image Features for Adaptive Retrieval

F. Carrara, G. Amato, F. Falchi, C. Gennaro. In International Conference on Multimedia Retrieval 2020, pp. 198-206. [14].

In the last years, content-based image retrieval largely benefited from representation extracted from deeper and more complex convolutional neural networks, which became more effective but also more computationally demanding. Despite existing hardware acceleration, query processing times may be easily saturated by deep feature extraction in high-throughput or real-time embedded scenarios, and usually, a trade-off between efficiency and effectiveness has to be accepted. In this work, we experiment with the recently proposed continuous neural networks defined by parametric ordinary differential equations, dubbed ODE-Nets, for adaptive extraction of image representations. Given the continuous evolution of the network hidden state, we propose to approximate the exact feature extraction by taking a previous "near-in-time" hidden state as features with a reduced computational cost. To understand the potential and the limits of this approach, we also evaluate an ODE-only architecture in which we minimize the number of classical layers in order to delegate most of the representation learning process — and thus the feature extraction process — to the continuous part of the model. Preliminary experiments on standard benchmarks show that we are able to dynamically control the trade-off between efficiency and effectiveness of feature extraction at inference-time by controlling the evolution of the continuous hidden state. Although ODE-only networks provide the best fine-grained control on the effectiveness-efficiency trade-off, we observed that mixed architectures perform better or comparably to standard residual nets in both the image classification and retrieval setups while using fewer parameters and retaining the controllability of the trade-off.

3.2.5
Learning Distance Estimators from Pivoted Embeddings of Metric Objects

F. Carrara, C. Gennaro, F. Falchi, G. Amato.
In the Proceedings of the 13th International Conference on Similarity Search and Applications (SISAP) 2020, pp 361-368. [15]

Efficient indexing and retrieval in generic metric spaces often translate into the search for approximate methods that can retrieve relevant samples to a query performing the least amount of distance computations. To this end, when indexing and fulfilling queries, distances are computed and stored only against a small set of reference points (also referred to as pivots) and then adopted in geometrical rules to estimate real distances and include or exclude elements from the result set. In this paper, we propose to learn a regression model that estimates the distance between a pair of metric objects starting from their distances to a set of reference objects. We explore architectural hyper-parameters and compare with the state-of-the-art geometrical method based on the n-simplex projection. Preliminary results show that our model provides a comparable or slightly degraded performance while being more efficient and applicable to generic metric spaces.

3.2.6
Unsupervised Vehicle Counting via Multiple Camera Domain Adaptation

L. Ciampi, C. Santiago, J.P. Costeira, C. Gennaro, G. Amato.
In Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI) 2020, CEUR Workshop Proceedings, pp. 82-85. [18]

Monitoring vehicle flows in cities is crucial to improve the urban environment and quality of life of citizens. Images are the best sensing modality to perceive and assess the flow of vehicles in large areas. Current technologies for vehicle counting in images hinge on large quantities of annotated data, preventing their scalability to city-scale as new cameras are added to the system. This is a recurrent problem when dealing with physical systems and a key research area in Machine Learning and AI. We propose and discuss a new methodology to design image-based vehicle density estimators with few labeled data via multiple camera domain adaptations.

3.2.7
Store Scientific Workflows Data in SSHOC Repository

C. Concordia, C. Meghini, F. Benedetti
Workshop about Language Resources for the SSH Cloud [23]

Today scientific workflows are used by scientists as a way to define automated, scalable, and portable in-silico experiments. Having a formal description of an experiment can improve replicability and reproducibility of the experiment. However, simply storing and publishing the workflow may be not enough, an accurate management of provenance data generated during workflow life cycle is crucial to achieve reproducibility. This document presents the activity being carried out by CNR-ISTI in task 5.2 of the SSHOC project to add to the repository service developed in the task, functionalities to store, access and manage 'workflow data' in order to improve replicability and reproducibility of e-science experiments.

3.2.8
L'Epistola a Cangrande al vaglio della Computational Authorship Verification: Risultati preliminari (con una postilla sulla cosiddetta "XIV Epistola di Dante Alighieri")

S. Corbara, A. Moreo, F. Sebastiani, M. Tavoni.

In Seminario "Nuove Inchieste sull'Epistola a Cangrande", Pisa University Press, 2020, pp. 153–192. [19].

In this work we apply techniques from computational Authorship Verification (AV) to the problem of detecting whether the "Epistle to Cangrande" is an authentic work by Dante Alighieri or is instead the work of a forger. The AV algorithm we use is based on "machine learning": the algorithm "trains" an automatic system (a "classifier") to detect whether a certain Latin text is Dante's or not Dante's, by exposing it to a corpus of example Latin texts by Dante and example Latin texts by authors coeval to Dante. The detection is based on the analysis of a set of stylometric features, i.e., style-related linguistic traits whose us- age frequencies tend to represent an author's unconscious "signature". The analysis carried out in this work suggests that, of the two parts into which the Epistle is traditionally subdivided, neither is Dante's. Experiments in which we have applied our AV system to each text in the corpus suggest that the system has a fairly high degree of accuracy, thus lending credibility to its hypothesis about the authorship of the Epistle. In the last section of this paper we apply our system to what has been hypothesized to be "Dante's 14th Epistle"; the system rejects, with very high confidence, the hypothesis that this epistle might be Dante's.

3.2.9
Edge-Based Video Surveillance with Embedded Devices

H. Kavalionak, C- Gennaro, G- Amato, C. Vairo, C. Perciante, C. Meghini, F. Falchi, and F. Rabitti

In 28th Italian Symposium on Advanced Database Systems, SEBD 2020, CEUR Workshop Proceedings, 2020, pp. 278–285. [27]

Video surveillance systems have become indispensable tools for the security and organization of public and private areas. In this work, we propose a novel distributed protocol for an edge-based face recognition system that takes advantage of the computational capabilities of the surveillance devices (i.e., cameras) to perform person recognition. The cameras fall back to a centralized server if their hardware capabilities are not enough to perform the recognition. We evaluate the proposed algorithm via extensive experiments on a freely available dataset. As a prototype of surveillance embedded devices, we have considered a Rasp- berry PI with the camera module. Using simulations, we show that our algorithm can reduce up to 50% of the load of the server with no negative impact on the quality of the surveillance service.

3.2.10
Cross-Resolution Deep Features Based Image Search

F.V. Massoli, F. Falchi, C.Gennaro, G. Amato

In International Conference on Similarity Search and Applications, SISAP 2020. [34]

Deep Learning models proved to be able to generate highly discriminative image descriptors, named deep features, suitable for similarity search tasks such as Person Re-Identification and Image Retrieval. Typically, these models are trained by employing high-resolution datasets, therefore reducing the reliability of the produced representations when low-resolution images are involved. The similarity search task becomes even more challenging in the cross-resolution scenarios, i.e., when a low-resolution query image has to be matched against a database containing descriptors generated from images at different, and usually high, resolutions. To solve this issue, we proposed a deep learning-based approach by which we empowered a ResNet-like architecture to generate resolution-robust deep features. Once trained, our models were able to generate image descriptors less brittle to resolution variations, thus being useful to fulfill a similarity search task in cross-resolution scenarios. To asses their performance, we used synthetic as well as natural low-resolution images. An immediate advantage of our approach is that there is no need for Super-Resolution techniques, thus avoiding the need to synthesize queries at higher resolutions.

3.2.11
kNN-guided Adversarial Attacks

F. V. Massoli, F. Falchi and G. Amato

In 28th Italian Symposium on Advanced Database Systems, SEBD 2020. [24]

In the last decade, we have witnessed a renaissance of Deep Learning models. Nowadays, they are widely used in industrial as well as scientific fields, and noticeably, these models reached super-human performances on specific tasks such as image classification. Unfortunately, despite their great success, it has been shown that they are vulnerable to adversarial attacks - images to which a specific amount of noise imperceptible to human eyes have been added to lead the model to a wrong decision. Typically, these malicious images are forged, pursuing a misclassification goal. However, when considering the task of Face Recognition (FR), this principle might not be enough to fool the system. Indeed, in the context FR, the deep models are generally used merely as features extractors while the final task of recognition is accomplished, for example, by similarity measurements. Thus, by crafting adversarials to fool the classifier, it might not be sufficient to fool the overall FR pipeline. Starting from this observation, we proposed to use a k-Nearest Neighbour algorithm as guidance to craft adversarial attacks against an FR system. In our study, we showed how this kind of attack could be more threatening for an FR system than misclassification-based ones considering both the targeted and untargeted attack strategies.

3.2.12
Heterogeneous document embeddings for cross-lingual text classification

A. Moreo, A. Pedrotti, F. Sebastiani. In 36th ACM Symposium on Applied Computing (SAC 2021). [43].

Funnelling *(*FUN*) is a method for cross-lingual text classification (CLC) based on a two-tier ensemble for heterogeneous transfer learning. In* FUN*, 1st-tier classifiers, each working on a different, language-dependent feature space, return a vector of calibrated posterior probabilities (with one dimension for each class) for each document, and the final classification decision is taken by a metaclassifier that uses this vector as its input. The metaclassifier can thus exploit class-class correlations, and this (among other things) gives* FUN *an edge over CLC systems where these correlations cannot be leveraged. We here describe* Generalized Funnelling *(*GFUN*), a learning ensemble where the metaclassifier receives as input the above vector of calibrated posterior probabilities, concatenated with document embeddings (aligned across languages) that embody other types of correlations, such as word-class correlations (as encoded by* Word-Class Embeddings*) and word-word correlations (as encoded by* Multilingual Unsupervised or Supervised Embeddings*). We show that* GFUN *improves on* FUN *by describing experiments on two large, standard multilingual datasets for multi-label text classification.*

3.2.13
Re-assessing the "classify and count" quantification method

A. Moreo and F. Sebastiani.
In Proceedings of the 43rd European Conference on Information Retrieval (ECIR 2021). [45]

Learning to quantify *(a.k.a.* quantification*) is a task concerned with training unbiased estimators of class prevalence via supervised learning. This task originated with the observation that "Classify and Count" (CC), the trivial method of obtaining class prevalence estimates, is often a biased estimator, and thus delivers suboptimal quantification accuracy; following this observation, several methods for learning to quantify have been proposed that have been shown to outperform CC. In this work we contend that previous works have failed to use properly optimised versions of CC. We thus reassess the real merits of CC (and its variants), and argue that, while still inferior to some cutting-edge methods, they deliver near-state-of-the-art accuracy once (a) hyperparameter optimisation is performed, and (b) this optimisation is performed by using a true quantification loss instead of a standard classification-based loss. Experiments on three publicly available binary sentiment classification datasets support these conclusions.*

3.2.14
Automatic Pass Annotation from Soccer VideoStreams Based on Object Detection and LSTM

D. Sorano, F. Carrara, P. Cintia, F. Falchi, L. Pappalardo.
In Proceedings of the European Conference on Machine Learning (ECML-PKDD), 2020. [50]

Soccer analytics is attracting increasing interest in academia and industry, thanks to the availability of data that describe all the spatio-temporal events that occur in each match. These events (e.g., passes, shots, fouls) are collected by human operators manually, constituting a considerable cost for data providers in terms of time and economic resources. In this paper, we describe PassNet, a method to recognize the most frequent events in soccer, i.e., passes, from video streams. Our model combines a set of artificial neural networks that perform feature extraction from video streams, object detection to identify the positions of the ball and the players, and classification of frame sequences as passes or not passes. We test PassNet on different scenarios, depending on the similarity of conditions to the match used for training. Our results show good classification results and significant improvement in the accuracy of pass detection with respect to baseline classifiers, even when the match's video conditions of the test and training sets are considerably different. PassNet is the first step towards an automated event annotation system that may break the time and the costs for event annotation, enabling data collections for minor and non-professional divisions, youth leagues and, in general, competitions whose matches are not currently annotated by data providers.

3.2.15
The Hypermedia Dante Network Project

G. Tomazzoli, L.M.G. Livraghi, D. Metilli, N. Pratelli, V. Bartalesi.

AIUCD 2021 Conference, 2020.[51]

3.3 Magazines

In this section, we report the paper we published in magazines during 2020 in alphabetic order of the first author.

3.3.1
Report on the 2nd ACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search

T. Berger-Wolf, B. Carterette, T. Elsayed, M. Keet, F. Sebastiani, H. Suleman.
In SIGIR Forum, vol. 54. [8].

We report on the organization and activities of the 2nd ACM SIGIR/SIGKDD Africa School on Machine Learning for Data Mining and Search, which took place at the University of Cape Town in South Africa January 27–31, 2020.

3.3.2
Transitioning the Information Retrieval Literature to a Fully Open Access Model

D. Hiemstra, M.-F. Moens, R. Perego, F. Sebastiani.
In SIGIR Forum, vol. 54. [26].

Almost all of the important literature on Information Retrieval (IR) is published in subscription-based journals and digital libraries. We argue that the lack of open access publishing in IR is seriously hampering progress and inclusiveness of the field. We propose that the IR community starts working on a road map for transitioning the IR literature to a fully, "diamond", open access model.

3.4 Preprints

3.4.1
The VISIONE Video Search System: Exploiting Off-the-Shelf Text Search Engines for Large-Scale Video Retrieval

Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, Claudio Vairo

arXiv:2008.02749. [2]

In this paper, we describe VISIONE, a video search system that allows users to search for videos using textual keywords, occurrence of objects and their spatial relationships, occurrence of colors and their spatial relationships, and image similarity. These modalities can be combined together to express complex queries and satisfy user needs. The peculiarity of our approach is that we encode all the information extracted from the keyframes, such as visual deep features, tags, color and object locations, using a convenient textual encoding indexed in a single text retrieval engine. This offers great flexibility when results corresponding to various parts of the query needs to be merged. We report an extensive analysis of the system retrieval performance, using the query logs generated during the Video Browser Showdown (VBS) 2019 competition. This allowed us to fine-tune the system by choosing the optimal parameters and strategies among the ones that we tested.

3.4.2
MedLatin1 and MedLatin2: Two Datasets for the Computational Authorship Analysis of Medieval Latin Texts

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, Mirko Tavoni.

arXiv 2011.08091. [20].

We present and make available MedLatin1 and MedLatin2, two datasets of medieval Latin texts to be used in research on computational authorship analysis. MedLatin1 and MedLatin2 consist of 294 and 30 curated texts, respectively, labelled by author, with MedLatin1 texts being of an epistolary nature and MedLatin2 texts consisting of literary comments and treatises about various subjects. As such, these two datasets lend themselves to supporting research in authorship analysis tasks, such as authorship attribution, authorship verification, or same-author verification.

3.4.3
Combining GANs and AutoEncoders for Efficient Anomaly Detection

Fabio Carrara, Giuseppe Amato, Luca Brombin, Fabrizio Falchi, Claudio Gennaro

arXiv:2011.08102 [13]. Accepted at the 25th International Conference on Patter Recognition (ICPR) 2020.

In this work, we propose CBiGAN — a novel method for anomaly detection in images, where a consistency constraint is introduced as a regularization term in both the encoder and decoder of a BiGAN. Our model exhibits fairly good modeling power and reconstruction consistency capability. We evaluate the proposed method on MVTec AD — a real-world benchmark for unsupervised anomaly detection on high-resolution images — and compare against standard baselines and state-of-the-art approaches. Experiments show that the proposed method improves the performance of BiGAN formulations by a large margin and performs comparably to expensive state-of-the-art iterative methods while reducing the computational cost. We also observe that our model is particularly effective in texture-type anomaly detection, as it sets a new state of the art in this category. Our code is available at `https://github.com/fabiocarrara/cbigan-ad`.

3.4.4
Training Convolutional Neural Networks with Hebbian Principal Component Analysis

G. Lagani, G. Amato, F. Falchi and C. Gennaro arXiv:2012.12229 [28]

Recent work has shown that biologically plausible Hebbian learning can be integrated with backpropagation learning (backprop), when training deep convolutional neural networks. In particular, it has been shown that Hebbian learning can be used for training the lower or the higher layers of a neural network. For instance, Hebbian learning is effective for re-training the higher layers of a pre-trained deep neural network, achieving comparable accuracy w.r.t. SGD, while requiring fewer training epochs, suggesting potential applications for transfer learning. In this paper we build on these results and we further improve Hebbian learning in these settings, by using a nonlinear Hebbian Principal Component Analysis (HPCA) learning rule, in place of the Hebbian Winner Takes All (HWTA) strategy used in previous work. We test this approach in the context of computer vision. In particular, the HPCA rule is used to train Convolutional Neural Networks in order to extract relevant features from the CIFAR-10 image dataset. The HPCA variant that we explore further improves the previous results, motivating further interest towards biologically plausible learning algorithms.

3.4.5
Assessing Pattern Recognition Performance of Neuronal Cultures through Accurate Simulation

G. Lagani, R. Mazziotti, F. Falchi, C. Gennaro, G.M. Cicchini, T. Pizzorusso, F. Cremisi, G. Amato arXiv:2012.10355 [29]

Previous work has shown that it is possible to train neuronal cultures on Multi-Electrode Arrays (MEAs), to recognize very simple patterns. However, this work was mainly focused to demonstrate that it is possible to induce plasticity in cultures, rather than performing a rigorous assessment of their pattern recognition performance. In this paper, we address this gap by developing a methodology that allows us to assess the performance of neuronal cultures on a learning task. Specifically, we propose a digital model of the real cultured neuronal networks; we identify biologically plausible simulation parameters that allow us to reliably reproduce the behavior of real cultures; we use the simulated culture to perform handwritten digit recognition and rigorously evaluate its performance; we also show that it is possible to find improved simulation parameters for the specific task, which can guide the creation of real cultures.

3.4.6
MOCCA: Multi-Layer One-Class Classification forAnomaly Detection

Fabio Valerio Massoli, Fabrizio Falchi, Alperen Kantarci, Şeymanur Akti,Hazim Kemal Ekenel, Giuseppe Amato arXiv:2012.12111 [35]

Anomalies are ubiquitous in all scientific fields and can express an unexpected event due to incomplete knowledge about the data distribution or an unknown process that suddenly comes into play and distorts the observations. Due to such events' rarity, it is common to train deep learning models on "normal", i.e. non-anomalous, datasets only, thus letting the neural network to model the distribution beneath the input data. In this context, we propose our deep

learning approach to the anomaly detection problem named Multi-Layer One-Class Classification (MOCCA). We explicitly leverage the piece-wise nature of deep neural networks by exploiting information extracted at different depths to detect abnormal data instances. We show how combining the representations extracted from multiple layers of a model leads to higher discrimination performance than typical approaches proposed in the literature that are based neural networks' final output only. We propose to train the model by minimizing the L_2 distance between the input representation and a reference point, the anomaly-free training data centroid, at each considered layer. We conduct extensive experiments on publicly available datasets for anomaly detection, namely CIFAR10, MVTec AD, and ShanghaiTech, considering both the single-image and video-based scenarios. We show that our method reaches superior performances compared to the state-of-the-art approaches available in the literature. Moreover, we provide a model analysis to give insight on how our approach works.

3.4.7
Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, Stéphane Marchand-Maillet. arXiv 2008.05231. [38]

Despite the evolution of deep-learning-based visual-textual processing systems, precise multi-modal matching remains a challenging task. In this work, we tackle the task of cross-media retrieval through image-sentence matching based on word-region alignments, using supervision only at the global image-sentence level. Specifically, we present a novel approach called Transformer Encoder Reasoning and Alignment Network (TERAN). TERAN enforces a fine-grained match between the underlying components of images and sentences, i.e., image regions and words, respectively, in order to preserve the informative richness of both modalities. TERAN obtains state-of-the-art results on the image retrieval task on both MS-COCO and Flickr30k datasets. Moreover, on MS-COCO, it also outperforms current approaches on the sentence retrieval task. Focusing on scalable cross-modal information retrieval, TERAN is designed to keep the visual and textual data pipelines well separated. Cross-attention links invalidate any chance to separately extract visual and textual features needed for the online search and the offline indexing steps in large-scale retrieval systems. In this respect, TERAN merges the information from the two domains only during the final alignment phase, immediately before the loss computation. We argue that the fine-grained alignments produced by TERAN pave the way towards the research for effective and efficient methods for large-scale cross-modal information retrieval. We compare the effectiveness of our approach against relevant state-of-the-art methods. On the MS-COCO 1K test set, we obtain an improvement of 5.7% and 3.5% respectively on the image and the sentence retrieval tasks on the Recall@1 metric.

3.4.8
Tweet sentiment quantification: An experimental re-evaluation

Alejandro Moreo and Fabrizio Sebastiani. arXiv 2011.08091. [44].

Sentiment quantification is the task of estimating the relative frequency (or "prevalence") of sentiment-related classes (such as Positive, Neutral, Negative) in a sample of unlabelled texts; this is especially important when these texts are tweets, since most sentiment classification endeavours carried out on Twitter data actually have quantification (and not the classification of individual tweets) as their ultimate goal. It is well-known that solving quantification via "classify and count" (i.e., by classifying all unlabelled items via a standard classifier and counting the items that have been assigned to a given class) is suboptimal in terms of accuracy, and that more accurate quantification methods exist. In 2016, Gao and Sebastiani carried out a systematic comparison of quantification methods on the task of tweet sentiment quantification. In hindsight, we observe that the experimental protocol followed in that work is flawed, and that its results are thus unreliable. We now re-evaluate those quantification methods on the very same datasets, this time following a now consolidated and much more robust experimental protocol, that involves 5775 as many experiments as run in the original study. Our experimentation yields results dramatically different from those obtained by Gao and Sebastiani, and thus provide a different, much more solid understanding of the relative strengths and weaknesses of different sentiment quantification methods.

4. Tutorials

4.1 Learning to Quantify

Alejandro Moreo and Fabrizio Sebastiani, "Learning to Quantify: Supervised Prevalence Estimation for Computational Social Science", half-day tutorial delivered at the *12th International Conference on Social Informatics (SocInfo 2020)*, Pisa, IT, October 2020.

5. Dissertations

5.1 MSc Dissertations

5.1.1
Development and Experimenting deep learning methods for unsupervised anomaly detection in images

Luca Brombin, MSc in Computer Engineering, University of Pisa, 2020 [9]. Advisors: Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato.

Anomaly detection in the industrial sector is an important problem as it is a key component of quality control systems that minimize the chance to miss a defective product. Most often, the anomaly detection is done through analysis of the images of the products. Because the products, or their designs change and quality data is hard to obtain, this problem is approached in an unsupervised manner. There are many different anomaly detection approaches, but most of them deal with low dimensional data and do not work well with the images. We examine deep learning techniques that utilize convolutional neural networks which can extract meaningful image representations to a lower-dimensional space. It allows the models to learn the important features of an image, regardless of some small changes in the input. The feature extracted with the CNN are used to train standard one class classifier (such as one class support vector

machine) that is able to classify an object in an unsupervised way. Then next approach is an anomaly detection based on generative adversarial networks (GAN). The network learns a mapping from the latent space to a representation of a "normal" data and is able to produce new and unseen data samples from random latent vectors. In particular, the main state of the art models for anomaly detection based on GAN were examined. Afterwards has been developed a new model (based on BiGAN) to detect anomalies to remove the problems of the standard methods and increase the performance called CBiGAN.

5.1.2
Design and implementation of attribute retrieval systems based on deep learning

Francesco Buiaroni, MSc in Computer Engineering, University of Pisa, 2020 [10]. Advisors: Claudio Gennaro, Fabio Valerio Massoli, Giuseppe Amato, Fabrizio Falchi.

Attribute-based image retrieval is a type of cross-modal retrieval system, in which data is described by two modalities, an image and an attribute, and the attribute is used as a query to return the image that satisfies it. It can be used in the field of surveillance to simplify the work of human personnel, returning images from a large database that meet certain attributes, without the human personnel having to check each image individually. To build the attribute retrieval system, we use approaches based on deep neural networks, which have the advantage of learning from data to perform a certain task. Specifically, convolutional neural networks (CNN) and multi-layer perceptrons (MLP) are used. In this work, we take into account two different scenarios: attribute retrieval on faces and attribute retrieval on vehicles. For attribute retrieval on faces and attribute retrieval on vehicles, we use an Attribute-based Deep Cross-Modal Hashing (ADCMH) framework, which is composed of two deep neural networks with different architecture. For vehicles only, in addition to ADCMH, two other approaches are tested. In the first approach, we test the ADCMH framework without quantization, i.e. removing the final hashing. The second approach is simpler and uses a single CNN trained as a multi-class classifier on vehicles to perform attribute retrieval.

5.1.3
Design and implementation of an efficient orbital debris detection in astronomical images using Deep Learning

Alessandro Cabras, MSc in Computer Engineering, University of Pisa, 2020 [11]. Advisors: Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato.

Wide-field telescopes that working in staring mode are widely used for optical astronomical observations. In the observed images, the identification of moving objects, visible as linear features (streaks), is important for several reasons including the cataloging of space debris. This thesis work consists in the design and development of a system based on machine learning approaches that identify these objects in real-time in the shortest possible time and return their position within the image through bounding boxes. In particular, the detection of the streak will be done in real-time during observation, using high-performance detection algorithms such as YOLO, Faster R-CNN, and SSD, in order to have time to be able to track the object with a second, reduced FOV, telescope. Machine learning models will be trained partly using synthetic images produced with a simulator provided by POLIMI and partly by real ones, kindly provided by the Italian Air Force's Experimental Flight Center. The results obtained will be useful for the realization of a tracking system that predicts the direction of movement of the debris and communicates it to a second telescope, with reduced FOV, which can track it.

5.1.4
Design and implementation of an application for art paintings classification and retrieval based on artificial intelligence

Roberto Magherini, MSc in Computer Engineering, University of Pisa, 2020 [30]. Advisors: Claudio Gennaro, Lucia Vadicamo, Giuseppe Amato, and Fabrizio Falchi

The purpose of this thesis is the development of a Web Application to categorize paintings and search for similarity by style. For this purpose, a Convolutional Neural Network (CNN) has been trained on two datasets, one of 13 style classes and one of 91 artist classes. Determining the style and artist of a painting can be difficult even for an expert, and sometimes it is also possible for two experts to express different opinions. Also, if it is possible to determine a unique artist for a painting, it is much more difficult to understand if the style is one or if it has not been influenced by other styles. The challenge is to be able to extract the true style of the painting and to be able to identify its artist regardless of the period in which she/he made it. Besides, another challenge is that the datasets available for this type of task are not large enough to allow the training of networks from scratch. In this thesis work, we used the following CNNs: VGG-16, VGG-19, ResNet50. The CNNs training is composed of two parts and the dataset Paintings91 coming from the University of Barcelona was used. The first part is based on the artist's classification. The CNNs were trained using the dataset organized according to the artist to whom the work belongs. The second part is based on style classification. The CNNs were trained using the dataset organized according to the style of the works. For both parts we used a process of transfer learning to reduce the training time and to be able to take advantage of a dataset not so big to allow the training from scratch. In particular, in order to achieve the best accuracy, a process of network tuning has been used. The framework used for training and testing of networks is Tensorflow, using the Keras API. To speed up the training and testing process, NVIDIA CUDA was exploited. The Keras API provides simple, fast, and efficient methods to use existing models, modify them, create new ones, and perform all the processes necessary to train and test neural networks. Once the best CNN was identified, a study was done on the classes that give the best and worst results to understand the cause of a good classification and a bad classification. In order to show the potential of these networks, an application was developed. It is a Web Application with a simple and intuitive main web page, where users can use an image as query and get information on the style and artist classification of the image. In addition, the Application performs a visual similarity search and provides users with the most similar images to the query image, which are obtained using a NoSQL database implemented through ElasticSearch. The application can be accessed from any device via the web and HTTP, since an HTTP Server has been im-

plemented using Flask. The HTTP server handles all user requests and interfaces directly with CNN and via REST calls with ElasticSearch. Based on the results of this research, we have obtained that the best accuracy is obtained through the use of residual networks (ResNet50). Therefore, this network was chosen for the development of the web Application.

5.1.5
Developing and Experimenting Approaches for DeepFake Text Detection on Social Media

Margherita Gambini, MSc in Computer Engineering, University of Pisa, 2020 [25]. Advisors: Maruzio Tesconi and Fabrizio Falchi.

5.1.6
Design and implementation of an anomaly detection system for videos

Edoardo Sassu, MSc in Computer Engineering, University of Pisa, 2020 [48]. Advisors: Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, and Claudio Gennaro

Anomaly detection consists in finding events or items which vary from the normality. It can be a useful tool to reduce or simplify the work that humans have to do, increasing productivity and reducing errors and costs. In this work, we take into account anomaly detection in videos. The identification of video anomalies does not require the use of specific sensors or equipments for a given scenario other than a camera, which makes visual anomaly detection very versatile and applicable to a wide range of scenarios. A suitable candidate to build an automatic anomaly detection system are Deep Convolutional Neural Networks (CNN) that have proven to be effective in Computer Vision tasks. The major feature of NNs is that they can learn from examples, with no need for any previous expertise or knowledge. This can be a useful feature in anomaly detection in which events could be unknown because of the sporadic nature of anomalies or challenging to represent. A viable approach is to train a model to learn how normality appears in an unsupervised way and considers all the events different from it as abnormal. The main advantage of this approach is that the system can be trained using an exhaustive dataset of normality examples. Most of the recent researches on anomaly detection, and also this work, go in this direction. This work aims to implement a frame prediction based anomaly detection system that performs as well as other state-of-the-art approaches and tests its discrimination capabilities on several types of anomalies. A custom dataset was also created to remedy the lack of some types of anomalies in publicly available datasets and tests the proposed solution with a wider range of anomalies.

5.1.7
Heterogeneous Document Embeddings for Multi-Lingual Text Classification

Andrea Pedrotti, MSc in Digital Humanities, University of Pisa, 2020 [46]. Advisors: Alejandro Moreo and Fabrizio Sebastiani.

Supervised Text Classification (TC) is a NLP task in which, given a set of training documents labelled according to a finite number of classes, a classifier is trained so that it maps unlabelled documents to the class or classes to which they are assumed to belong, based on the document's content. For a classifier to be trained, documents need first to be turned into vectorial representations. While this has been traditionally achieved utilizing the BOW ("bag of words") approach, the current research trend is to learn continuous and dense such representations, called embeddings*. Multi-lingual Text Classification (MLTC) is a specific setting of TC. In MLTC each document* $\mathbf{x}$ *is written in one of a finite set* $\mathscr{L} = \{\lambda_1, ..., \lambda_{|\mathscr{L}|}\}$ *of languages, and unlabelled documents need to be classified according to a common codeframe (or "classification scheme")* $\mathscr{C} = \{c_1, ..., c_{|\mathscr{C}|}\}$*. We approach MLTC by using funnelling, an algorithm originally proposed by Esuli et al. Funnelling is a two-tier ensemble-learning method, where the first tier trains language-dependent classifiers that generate document representations consisting of their posterior probabilities for the classes in the codeframe, and where the second tier trains a meta-classifier using all the (language-independent) probabilistic representations. In this thesis we redesign funnelling by generalizing this procedure; we call the resulting framework Generalized Funnelling (gFun). In doing so, we enable gFun's meta-classifier to capitalize on different language-independent views of the document, that go beyond the document-class correlations captured by the posterior probabilities that are used in "standard" funnelling. To exemplify such views, we experiment with embeddings derived from word-word correlations (for this we use MUSE embeddings) and embeddings derived from word-class correlations (for this we use "word-class embeddings") aligned across languages. The extensive empirical evaluation we have carried out seems indeed to confirm the hypothesis that multiple, language-independent views that capture different types of correlations are beneficial for MLTC.*

6. Datasets

Authorship Analysis of Medieval Latin

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. "Two Datasets for the Computational Authorship Analysis of Medieval Latin Texts."

We make available MedLatin1 and MedLatin2, two datasets of medieval Latin texts to be used in research on computational authorship analysis. MedLatin1 and MedLatin2 consist of 294 and 30 curated texts, respectively, labelled by author, with MedLatin1 texts being of an epistolary nature and MedLatin2 texts consisting of literary comments and treatises about various subjects. As such, these two datasets lend themselves to supporting research in authorship analysis tasks, such as authorship attribution, authorship verification, or same-author verification.

`https://doi.org/10.5281/zenodo.3903296`

7. Code

7.0.1
An authorship verification tool

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. "MedieValla: An authorship verification tool written in Python for medieval Latin."

`https://doi.org/10.5281/zenodo.3903236`

7.0.2
Transformer Encoder Reasoning Networks for Visual Textual Retrieval

Code for replicating the visual-textual retrieval experiments in [39, 38], written in Python with the use of the PyTorch framework. Transformer Encoder Reasoning Network (TERN):

`https://github.com/mesnico/TERN`

Transformer Encoder Reasoning and Alignment Network (TERAN):

`https://github.com/mesnico/TERAN`

7.0.3
Virtual to Real Pedestrian Detection

Code for replicating the experiments in [17]. The provided code trains the Faster R-CNN detector exploiting *ViPeD*, a synthetic collection of images suitable for the pedestrian detection task, and employing some Domain Adaptation techniques to tackle the *Synthetic2Real* Domain Shift.

`https://github.com/ciampluca/Virtual-to-Real-Pedestrian-Detection`

References

[1] G. Amato, F. Carrara, F. Falchi, C. Gennaro, F. Rabitti, and L. Vadicamo. Scalar quantization-based text encoding for large scale image retrieval. In *CEUR Workshop Proceedings*, volume 2646, pages 258–265, 2020.

[2] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. The visione video search system: Exploiting off-the-shelf text search engines for large-scale video retrieval, 2020.

[3] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. Large-scale instance-level image retrieval. *Information Processing & Management*, 57(6):102100, 2020.

[4] Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, Fabio Valerio Massoli, and Claudio Vairo. Multi-resolution face recognition with drones. In *to appear in 3rd International Conference on Sensors, Signal and Image Processing (SSIP)*, pages 1–8. ACM, 2020.

[5] Alessandro Artusi, Francesco Banterle, Fabio Carrara, and Alejandro Moreo. Efficient evaluation of image quality via deep-learning approximation of perceptual metrics. *IEEE Transactions on Image Processing*, 29:1843–1855, 2020.

[6] Francesco Banterle, Alessandro Artusi, Alejandro Moreo, and Fabio Carrara. Nor-vdpnet: A no-reference high dynamic range quality metric trained on hdr-vdp 2. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 126–130, 2020.

[7] Marco Di Benedetto, Fabio Carrara, Enrico Meloni, G. Amato, F. Falchi, and C. Gennaro. Learning accurate personal protective equipment detection from virtual worlds. *Multimedia Tools and Applications*, 2020.

[8] Tanya Berger-Wolf, Ben Carterette, Tamer Elsayed, C. Maria Keet, Fabrizio Sebastiani, and Hussein Suleman. Report on the 2nd ACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search. *SIGIR Forum*, 54(1), 2020.

[9] Luca Brombin. Development and experimenting deep learning methods for unsupervised anomaly detection in images. Master's thesis, MSc in Computer Engineering, University of Pisa, Italy, 2020.

[10] Francesco Buiaroni. Design and implementation of attribute retrieval systems based on deep learning. Master's thesis, MSc in Computer Engineering, University of Pisa, Italy, 2020.

[11] Alessandro Cabras. Design and implementation of an efficient orbital debris detection in astronomical images using deep learning. Master's thesis, MSc in Computer Engineering, University of Pisa, Italy, 2020.

[12] F. Carrara, R. Caldelli, F. Falchi, and G. Amato. On the robustness to adversarial examples of neural ode image classifiers. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2020.

[13] Fabio Carrara, Giuseppe Amato, Luca Brombin, Fabrizio Falchi, and Claudio Gennaro. Combining gans and autoencoders for efficient anomaly detection, 2020.

[14] Fabio Carrara, Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. Continuous ode-defined image features for adaptive retrieval. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 198–206, New York, NY, USA, 2020. Association for Computing Machinery.

[15] Fabio Carrara, Claudio Gennaro, Fabrizio Falchi, and Giuseppe Amato. Learning distance estimators from pivoted embeddings of metric objects. In *International Conference on Similarity Search and Applications*, pages 361–368. Springer, Cham, 2020.

[16] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

[17] Luca Ciampi, Nicola Messina, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Virtual to real adaptation of pedestrian detectors. *Sensors*, 20(18):5250, 2020.

[18] Luca Ciampi, Carlos Santiago, João Paulo Costeira, Claudio Gennaro, and Giuseppe Amato. Unsupervised vehicle counting via multiple camera domain adaptation. In Alessandro Saffiotti, Luciano Serafini, and Paul Lukowicz, editors, *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostella, Spain, September 4, 2020*, volume 2659 of *CEUR Workshop Proceedings*, pages 82–85. CEUR-WS.org, 2020.

[19] Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. L'epistola a Cangrande al vaglio della computational authorship verification: Risultati preliminari (con una postilla sulla cosiddetta "XIV Epistola di Dante Alighieri"). In Alberto Casadei, editor, *Atti del Seminario "Nuove Inchieste sull'Epistola a Cangrande"*, pages 153–192, Pisa, IT, 2020. Pisa University Press.

[20] Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. MedLatin1 and MedLatin2: Two datasets for the computational authorship analysis of medieval Latin texts, 2020. arXiv 2006.12289.

[21] Andrea Esuli, Alessio Molinari, and Fabrizio Sebastiani. A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment. *ACM Transactions on Information Systems*, 19(2):Article 19, 2020.

[22] Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. Cross-lingual sentiment quantification. *IEEE Intelligent Systems*, 35(3):106–114, 2020.

[23] Concordia C.; Meghini C.; Benedetti F. Store scientific workflows data in sshoc repository. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 1–4, Paris, 2020. European language resources association (ELRA).

[24] Fabrizio Falchi Fabio Valerio Massoli and Giuseppe Amato. knn-guided adversarial attacks. In *28th Italian Symposium on Advanced Database Systems (SEBD)*, 2020.

[25] Margherita Gambini. Developing and experimenting approaches for deepfake text detection on social media. Master's thesis, MSc in Computer Engineering, University of Pisa, Italy, 2020.

[26] Djoerd Hiemstra, Marie-Francine Moens, Raffaele Perego, and Fabrizio Sebastiani. Transitioning the information retrieval literature to a fully open access model. *SIGIR Forum*, 54(1), 2020.

[27] Hanna Kavalionak, Claudio Gennaro, Giuseppe Amato, Claudio Vairo, Costantino Perciante, Carlo Meghini, Fabrizio Falchi, and Fausto Rabitti. Edge-based video surveillance with embedded devices. In *in 28th Italian Symposium on Advanced Database Systems (SEBD)*, pages 278–285, 2020.

[28] Gabriele Lagani, Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. Training convolutional neural networks with hebbian principal component analysis, 2020.

[29] Gabriele Lagani, Raffaele Mazziotti, Fabrizio Falchi, Claudio Gennaro, Guido Marco Cicchini, Tommaso Pizzorusso, Federico Cremisi, and Giuseppe Amato. Assessing pattern recognition performance of neuronal cultures through accurate simulation, 2020.

[30] Roberto Magherini. Design and implementation of an application for art paintings classification and retrieval based on artificial intelligence. Master's thesis, MSc in Computer Engineering, University of Pisa, Italy, 2020.

[31] Fabio Valerio Massoli, Giuseppe Amato, and Fabrizio Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99:103927, 2020.

[32] Fabio Valerio Massoli, Fabio Carrara, Giuseppe Amato, and Fabrizio Falchi. Detection of face recognition adversarial attacks. *Computer Vision and Image Understanding*, 202:103103, 2020.

[33] Fabio Valerio Massoli, Fabrizio Falchi, and Giuseppe Amato. Cross-resolution face recognition adversarial attacks. *Pattern Recognition Letters*, 140:222 – 229, 2020.

[34] Fabio Valerio Massoli, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Cross-resolution deep features based image search. In *International Conference on Similarity Search and Applications*, pages 352–360. Springer, 2020.

[35] Fabio Valerio Massoli, Fabrizio Falchi, Alperen Kantarci, Şeymanur Akti, Hazim Kemal Ekenel, and Giuseppe Amato. Mocca: Multi-layer one-class classification foranomaly detection, 2020. arXiv:2012.12111.

[36] Metilli D. Meghini C., Bartalesi V. Introducing narratives in europeana: A case study. *Semantic Web*, 2020.

[37] Nicola Messina. Relational visual-textual information retrieval. In *International Conference on Similarity Search and Applications*, pages 405–411. Springer, 2020.

[38] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *arXiv preprint arXiv:2008.05231*, 2020.

[39] Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. Transformer reasoning network for image-text matching and retrieval. In *International Conference on Pattern Recognition (ICPR) 2020 (Accepted)*, 2020.

[40] Tomas Mikolov, Wen-Tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013)*, pages 746–751, Atlanta, US, 2013.

[41] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. Learning to weight for text classification. *IEEE Transactions on Knowledge and Data Engineering*, 32(2):302–316, 2020.

[42] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. Word-class embeddings for multiclass text classification. *Data Mining and Knowledge Discovery*, 2021. Forthcoming.

[43] Alejandro Moreo, Andrea Pedrotti, and Fabrizio Sebastiani. Heterogeneous document embeddings for cross-lingual text classification. In *Proceedings of the 36th ACM Symposium on Applied Computing (SAC 2021)*, Gwangju, KR, 2021. Forthcoming.

[44] Alejandro Moreo and Fabrizio Sebastiani. Tweet sentiment quantification: An experimental re-evaluation, 2020. arXiv 2011.08091.

[45] Alejandro Moreo and Fabrizio Sebastiani. Re-assessing the "classify and count" quantification method. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR 2021)*, Lucca, IT, 2021. Forthcoming.

[46] Andrea Pedrotti. Heterogeneous document embeddings for multi-lingual text classification. Master's thesis, MSc in Digital Humanities, University of Pisa, 2020.

[47] Lucas May Petry, Camila Leite Da Silva, Andrea Esuli, Chiara Renso, and Vania Bogorny. Marc: a robust method for multiple-aspect trajectory classification via space, time, and semantic embeddings. *International Journal of Geographical Information Science*, 34(7):1428–1450, 2020.

[48] Edoardo Sassu. Design and implementation of an anomaly detection system for videos. Master's thesis, MSc in Computer Engineering, University of Pisa, Italy, 2020.

[49] Fabrizio Sebastiani. Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal*, 23(3):255–288, 2020.

[50] Danilo Sorano, Fabio Carrara, Paolo Cintia, Fabrizio Falchi, and Luca Pappalardo. Automatic pass annotation from soccer videostreams based on object detection and lstm, 2020.

[51] Metilli D. Pratelli N. Bartalesi V. Tomazzoli G., Livraghi L. The hypermedia dante network project. In *Proceedings of the X Annual Conference of AIUCD*, 2021. Under publication.

[52] Lucia Vadicamo, Claudio Gennaro, Fabrizio Falchi, Edgar Chávez, Richard Connor, and Giuseppe Amato. Re-ranking via local embeddings: A use case with permutation-based indexing and the nsimplex projection. *Information Systems*, 95:101506, 2021.

[53] Xenophon Zabulis, Carlo Meghini, Nikolaos Partarakis, Cynthia Beisswenger, Arnaud Dubois, Maria Fasoula, Vito Nitti, Stavroula Ntoa, Ilia Adami, Antonios Chatziantoniou, et al. Representation and preservation of heritage crafts. *Sustainability*, 12(4):1461, 2020.