The final version of this manuscript appeared in the Proceedings of the International Conference on Computing, Networking and Communications (ICNC 2016) Kauai, Hawaii, USA, February 15-18, 2016
DOI: not yet available

COBANETS: a new paradigm for cognitive communications systems

Michele Zorzi, Andrea Zanella, Alberto Testolin, Michele De Filippo De Grazia, and Marco Zorzi

***Abstract*—In response to the new challenges in the design and operation of communication networks, and taking inspiration from how living beings deal with complexity and scalability, in this position paper we introduce an innovative system concept called COgnition-BAsed NETworkS (COBANETS). The proposed approach develops around the systematic application of advanced machine learning techniques and, in particular, *unsupervised deep learning and probabilistic generative models* for system-wide learning, modeling, optimization, and data representation. Moreover, in COBANETS we propose to combine the learning architecture with the emerging *network virtualization* paradigms, which make it possible to actuate automatic optimization and reconfiguration strategies at the system level, thus fully unleashing the potential of the learning approach.**

Compared to past and current research efforts in this area, the technical approach depicted in this paper is deeply interdisciplinary and more comprehensive, calling for the synergic combination of expertise of computer scientists, communications and networking engineers, and cognitive scientists, with the ultimate aim of breaking new ground through a profound rethinking of how the modern understanding of cognition can be used in the management and optimization of telecommunication networks.

I. Introduction

TRADITIONALLY, the ISO/OSI system architecture has been the cornerstone of network design, due to its modularity that enables the optimization of individual sets of functionalities and guarantees scalability. While such an ordered and simple structure has successfully served the needs of the Internet users up to now, the always increasing number and variety of services deployed over the network, and the effort of the Internet service providers to continuously improve the quality of the services offered to their customers are challenging the current network architecture, that suffers from ossification in the underlying infrastructure and does not appear capable of scaling up with the growing complexity of the upcoming communication scenarios.

This trend is indeed expected to accelerate in future fifth-generation (5G) mobile systems that, though not yet fully specified, will certainly pose extreme challenges in terms of *heterogeneity* of both device capabilities and traffic; *scalability* in terms of number of functions and parameters within a single node, and of number of nodes in the system; *efficient use of the resources*, such as bandwidth and energy; and effective management of *Quality of Experience* (QoE) [1]–[3].

Consider, for example, a *massive access* scenario, where the base stations are required to guarantee access to a very large number of machine-type devices that sporadically transmit very short packets [4], [5], or sustain the simultaneous upload of many pictures or videos taken by people attending a common public event. With current technologies and protocols, massive access will generate overwhelming signaling overhead and cause service outages. Another demanding scenario is that of *Heterogeneous Networks* (HetNets), in which many pico and/or femto base stations will be placed within a macro cell to provide better coverage and higher connection speed. However, the design of efficient handover policies, resource allocation/reservation schemes, service migration strategies, and so on in these systems are complex open problems that require new approaches and methodologies [6], [7]. A third challenging problem is to provide adequate Quality of Experience (QoE) to mobile users, irrespective of the number and variety of data flows that need to be simultaneously served by the wireless access network. To reach this objective, the new generations of communication systems shall be able to differentiate the services not only by class of application, but even per flow within each class, thus providing *content-based service optimization* [8], [9].

There is hence a need to manage more efficiently the available resources, taking into account the vast variety of traffic features and of their performance requirements, as well as the extreme heterogeneity of device capabilities and of communications technologies.

Recent trends in networking have shown that crossing the boundaries of the layering architecture can lead to much higher efficiency than respecting the orthodox layered model, and cross-layer approaches have been proposed and shown to provide very good results, especially in resource-challenged environments. Furthermore, newly emerged networking paradigms and capabilities, including Software Defined Networking (SDN) and Network Function Virtualization (NFV) [10], open up unprecedented opportunities towards new systems and applications.

However, greater flexibility in network management implies more degrees of freedom in the setting of the network parameters and, consequently, a much bigger optimization space, which will call for more advanced (and complex) optimization strategies. If in addition we try to use the abundance of sensory data already present (or easily obtainable) in networks and devices, the dimensionality of the problem quickly becomes very large, making traditional approaches insufficient and calling for disruptive paradigms.

As a response to these challenges, and inspired by how the nervous system of living beings deals with complexity and scalability, we introduce the new concept of COgnition-BAsed NETworkS (COBANETS), intelligent communications

Michele Zorzi and Andrea Zanella are with the Department of Information Engineering while Alberto Testolin, Michele D.F. De Grazia, and Marco Zorzi are with the Department of General Psychology of the University of Padova, PD, 35131, Italy. e-mail: {firstname.lastname}@unipd.it

systems which are much more than just a collection of smart or cognitive nodes, and instead include a network-wide cognitive infrastructure for learning, modeling and optimization, and data representation. Advanced machine learning techniques, in particular unsupervised deep learning and probabilistic generative models (suitable for scenarios with massive unlabeled data), along with network optimization at all layers of the protocol stack and corresponding reconfiguration through SDN tools, are the key building blocks of our approach, which significantly departs from state-of-the-art solutions in cognitive networking.

The conceptual design and practical implementation of cognition-based networks has been elusive for years. In this paper, we advocate that this vision is now at hand, because of the following key enabling factors:

i) the recent advances in cognitive science, with the development of deep unsupervised learning networks which have been successfully applied to solve extremely difficult classification problems;
ii) the impressive performance improvement of processing units, with the commercial diffusion of parallel computing architectures that are particularly suitable for running very-large-scale deep learning models;
iii) the rapidly growing popularity of new networking paradigms, such us SDN and NFV, that have the potential to overcome the ossification in the underlying infrastructure of the Internet and enable a more dynamic and flexible management of the network, thus making it possible to actuate network-wide optimization strategies.

Despite the conjunction of these favorable factors, building the grand vision of a learning network, able to adapt to changing conditions and to serve multiple communication services, still remains a great challenge, which requires pushing the research significantly beyond the current state-of-the-art.

In the rest of this position paper, we describe our vision on how to move forward towards the practical realization of the COBANETS concept. The paper is organized as follows. In Sec. II we will quickly survey the recent history of cognitive networking and of machine learning applied to network optimization. Furthermore, the section offers a brief introduction of unsupervised learning techniques, which are at the core of the COBANETS framework. The reasons behind this choice are discussed in Sec. III, which describes in more details the COBANETS concept, and the specific properties of Generative Deep Neural Networks that make them particularly attractive as key enabling elements of a cognitive architecture. Successively, Sec. IV discusses the most relevant research challenges opened by the proposed approach, and finally Sec. V concludes the paper with a short summary of the study and some final considerations.

II. State of the art

In order to set the stage for the description of the proposed approach, we first provide a brief overview of the recent history of cognitive networking and machine learning approaches, with a particular focus on deep learning and generative models which are the basic building blocks of our approach.

a) Cognitive radios and networks: Cognition as a way to deal with the challenges of future networks has been suggested several times in the past. The pioneering work in [11], [12] proposed to apply cognition to special communications devices, called cognitive radios, able to learn and adapt to the environment, with the goal of providing reliable communication and efficient utilization of the radio spectrum. This concept of adaptability at the physical layer was later extended to a paradigm called cognitive radio network [13], where the spectrum owned by the so-called primary users (i.e., the legitimate users of a licensed band) is shared by secondary cognitive radios (frequency-agile transceivers with enough intelligence to perform spectrum sensing and dynamic spectrum access, and to communicate while coexisting with the primary users).

Even though in most of the existing papers on cognitive radio and cognitive radio networks the "cognitive" aspects are focused on sensing, channel selection, and adaptive communications, both Mitola [12] and Haykin [11] actually gave a broader definition of cognitive radio, which includes aspects that relate to the true essence of cognition, such as intelligent observation, learning, and decision-making. From this viewpoint, the existing studies that have addressed these cognition issues in networks, though certainly interesting and valuable in their own right, have only scratched the surface of what promises to be a rich research area with high potential for innovation. In this direction, cognitive networks for wireless systems [14], [15] and the Knowledge Plane for the Internet [17] have been proposed as new paradigms in which the concepts of cognition, learning and adaptability are applied in an end-to-end fashion to the whole protocol stack.

b) The role of machine learning: Machine Learning (ML) tools have been recently used in a networking context (e.g., see [18] for a recent survey on learning techniques for cognitive radio networks). Some examples of topics that have been addressed in this context include architectural models [19], data routing and clustering in sensor and ad hoc networks [20], and optimization of routing and scheduling through reinforcement learning. Supervised learning methods have been widely used to address various classification tasks., e.g., IP packet and Internet traffic classification [21]. A computer program able to automatically design end-to-end congestion-control algorithms is proposed in [22] and proved to outperform the best-known techniques in many different scenarios, showing that an intelligent unified framework can do better than customized solutions.

c) Unsupervised learning and generative models: The above applications of ML to networking problems are meant to solve specific issues, and make use of either supervised learning (in which correct inputs/outputs are explicitly presented and/or suboptimal actions are explicitly corrected) or some form of reinforcement learning (where an agent receives a reward based on the action it chooses, trying to find a balance between exploration and exploitation). These approaches are typically effective in the presence of a well-defined goal and of a feedback loop through which the agent is informed about the goodness of a certain action. However, although humans often learn through supervision (by teacher instruction) or

reinforcement (by understanding the effects of actions), we also continuously perform unsupervised learning, in which through stimuli from the environment we gradually develop a worldview upon which we build our cognitive activities [23]. Unsupervised learning gives an agent the ability to deal with situations never encountered before, and can exploit the huge amount of unlabeled data to build rich internal representations, on which supervised tasks can be more easily carried out [24], a framework often associated with the notions of representation learning and transfer learning, where knowledge abstracted from one domain is readily re-used in many different supervised tasks.

In this context, a generative model is a probabilistic model of how the underlying physical properties of the world cause sensory data [25], that can be efficiently implemented in stochastic recurrent neural networks, such as the Boltzmann machine, recently formalized following [26]. These advances, and the efficient unsupervised training on big datasets, have made it possible for the very first time to effectively stack together several basic modules in order to learn hierarchical architectures [27], which paved the way for the so-called deep learning models [28]. Although most recent research on deep learning has focused on the use of supervised techniques, unsupervised deep learning remains the only choice when data cannot be easily labeled (as in our scenarios of interest), and represents a research frontier for the future [29]. Moreover, the introduction of powerful parallel computing architectures, such as the CUDA framework, now makes it possible to efficiently build very-large-scale deep learning models. By exploiting these advances, deep learning algorithms have recently led to impressive performance gains in many difficult ML tasks, e.g., object recognition, natural language modeling, and studying the effects of mutations in DNA, just to name a few. Generative neural networks have also been recently extended to model sequential data and have been combined with reinforcement learning strategies, approaching human-level performance [30].

III. The COBANETS concept

From the above review of the related literature, we learn that, although the powerful paradigm of bringing cognitive processes into networks has been suggested in various forms in the past fifteen years or so, the idea has not yet found its way into a comprehensive and practical design, and even less so to a large-scale application in real systems. We believe the main reasons for this are to be found in the lack of a sufficiently general tool to implement intelligence in a scalable way, and the lack of actionable schemes able to effectively implement decisions in complex systems, possibly combined with the lack of a broader view of the cognition-based system beyond the ad hoc application of specific ML techniques to a limited set of functionalities.

New paradigms that have emerged only very recently in the areas of cognitive science (deep networks and generative models, i.e., the intelligence) and networking (software defined networks, i.e., the actionable schemes) make this the right time for a disruptive change of paradigm and for realizing

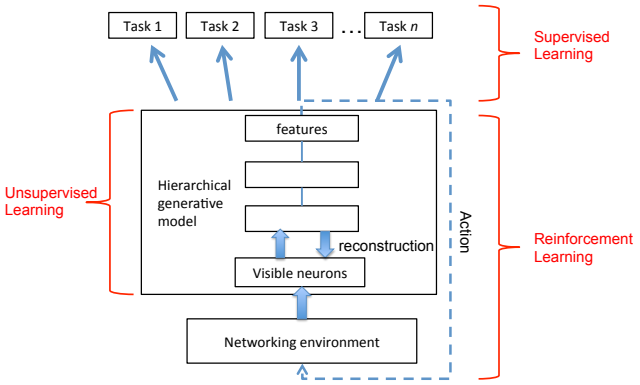

Fig. 1. Schematic representation of a learning framework based on a Generative Deep Neural Network (GDNN)

the ambition to bring cognitive networking techniques to the next level, by moving from the limited scope of a set of specific applications towards the development of a comprehensive framework in which large-scale unsupervised learning is the stepping stone and a key enabler of a wide range of optimization techniques for the whole network, as well as for its individual components.

According to these premises, the COBANETS concept focuses on Generative Deep Neural Networks (GDNN) and network virtualization paradigms as key enabling factors for the development of a groundbreaking novel approach to network optimization. In the remainder of this section, we describe in more details the characteristics that make GDNN extremely appealing in this context and, then, we give a broad description of the system architecture we envision.

A. GDNN: Generative Deep Neural Networks

At an intuitive level, the unsupervised training of a deep neural network builds an inner model of unlabeled input signals that is independent of any specific concept defined by the user. Unlike in typical supervised learning tasks, here the learning objective is to extract a useful set of features from the input space, which allow to accurately represent and reconstruct the input through a specific configuration of the neurons in the deeper layer (features). The generative model obtained from unsupervised training of a deep neural network (hereafter called Generative Deep Neural Network GDNN) supports the general learning framework represented in Fig. 1, and is characterized by the following specific properties.

Generative property: GDNNs are trained to minimize the error between the observed data and its estimate obtained from the inner representation of the input, given by the hidden layer. This property can be used for predictions and anomaly detection.

Feature extraction: The internal representations extracted by unsupervised learning are not tied to a specific discriminative task and turn out to be generally more informative than those obtained with supervised training. This property can be used to enhance the performance of simple supervised techniques applied to the features instead of the original data [31], [32].

Compact data representation: Deep unsupervised learning can also be interpreted as a particular type of efficient coding

strategy. This property can be used for data compression and dimensionality reduction, and to achieve scalability [27].

Synergy with reinforcement learning: The generative approach offers new insights about the possible role of reinforcement learning, by which an agent can improve its own internal model of the world by actively searching for information that can be used to disambiguate competing hypotheses [30], [33].

These properties of GDNNs can be exploited to develop a system architecture capable of efficiently dealing with the scalability, management, and multipurpose optimization challenges offered by the next generations of communication systems and services. For example, GDNNs can be trained to learn the wireless channel model for a mobile user and then exploited to predict the evolution of the wireless channel and proactively adapt all protocol layers accordingly. Similarly, a GDNN may be used to infer the nature of the traffic generated by the devices, thus making it possible to discriminate not only between different classes of traffic sources (e.g, alarms versus periodic metering data), but even between different streams of the same class (e.g., between sensory data with higher or lower priority depending, for instance, on their temporal trends). A practical example of the way GDNNs can be exploited to gain better context information and optimize QoE is given in our recent work [34], where we trained a GDDN to learn a generative model of the *size* (not the specific content) of encoded video frames. This model was then used to estimate the rate/distortion curve of each video sequence and, then, to design QoE-aware resource allocation and call admission control algorithms [35], [36].

B. General architecture

To fully express the potential of GDNNs, we envision an architecture that enables network-wide observation and sensing at multiple levels, including quantities such as protocol parameters and state variables, traffic conditions, channel statistics, transmission and error events, interference, and so on. The architecture shall also provide the flexibility required for the practical implementation of the proposed approach. For this reason, we look at the SDN and NFV paradigms as basic building blocks of COBANETS. An advanced SDN controller, indeed, may be able to collect the inter-device data generated by the cognitive nodes in the system, and make decisions for system-wide optimization. SDN protocols, such as OpenFlow, can be used to implement the optimization actions on the different nodes, whose functionalities shall be completely virtualized per the NFV concept. In addition, the controller may instruct the nodes to instantiate generative deep learning modules for the optimization of local functionalities (e.g., PHY and MAC), using inter-device data only.

Clearly, there is still a long way to go to turn this broad and very general vision into a practical and well defined system architecture. In the remainder of this section, we describe some of the main components that we believe may help reach this ultimate objective. These components, which shall all be part of the final COBANETS architecture, are here presented in order of increasing generality and scope, thus reflecting the natural path we envision for the development of the COBANETS concept.

1) Functional abstraction and optimization: Important components of COBANETS will be generative models that provide an informative representation of fundamental elements and functionalities of a communication network, including traffic sources, radio channel, MAC protocols, and so on. In parallel, we may have GDNNs that can capture the interdependencies among the parameters within a certain protocol layer (e.g., at the physical layer the transmission parameters of a mobile node and the interference from adjacent cells, or at the MAC layer the packet inter-arrival time and the number of retransmissions). These generative models can be used to predict the offered traffic in the near future and/or to train classifiers to get more detailed context information, for instance the type of application(s) generating the data flows, the operational scenarios (indoor, urban, vehicular, rural), or the congestion level of a certain connection. The context information, in turn, may be used to optimize some network functionalities (e.g., handover, content caching, transmit rate, and so on).

2) Integration of different generative models: In analogy with the sensory segregation and integration observed in the brain, the specialized modules operating in different domains, as above described, should be combined in a learning architecture capable of building more abstract representations of the world. Implementing this strategy is very challenging, because it requires to carefully engineer the scope of each sub-module and to integrate the internal representations created by different models without disrupting domain-specific knowledge. A possible solution can be to concatenate the representation of different models and to jointly train an additional generative model with the task of reconstructing this composite input, thereby learning useful correlations among the abstract representations provided by different sensory domains [37]. Another approach consists in mapping the abstract representations learned by the sub-models into layer-specific performance indices and then training a combined generative model using such indices. These two approaches, as well as others, shall be deeply studied and compared in terms of complexity and efficiency, with the goal of identifying the best solution.

3) Inter flow optimization: As an intermediate step to system-wide optimization, we believe that COBANETS shall make it possible to jointly optimize multiple functionalities with local scope (e.g., within a single node). Unlike in traditional cross-layer optimization, where prior knowledge of some explicit interdependencies among protocols is assumed, the approach based on generative model learning has the potential to discover and exploit hidden relations among the different parameters, which can be specific for a certain application scenario or user profile and, hence, are not replicable in other contexts. For example, daily habits of users (e.g., watching movie trailers on the smartphone while commuting by train) may be reflected in specific inter-relations among the type of traffic generated by the device, the interference produced by other devices, and the radio channel characteristics. Therefore, COBANETS shall entail generative models capable of capturing these multifold correlations, thus supporting the design of optimization strategies that are adapted to a specific

device and scenario.

4) System Level optimization: The final objective of COBANETS shall have a global scope, and shall refer to the whole system, rather than to single nodes or flows. A possible global optimization may regard the routing strategy, scheduling policies in the switches, resource allocation at base stations, and transport protocol parameters, which can be jointly optimized according to the nature of the data, the characteristics of the user (static, mobile), the congestion on the links, and so on. We believe that GDNNs can enable the development of an innovative scalable approach to the above problem, by taking advantage of the data provided by the single agents of the system and collected by the cognition-based architecture designed, thus making it possible for a centralized network controller to autonomously derive strategies for the maximization of multi-objective functions, and to actuate such strategies in the network elements by means, e.g., of SDN.

IV. Research challenges

Despite the encouraging results of some initial studies, the design of an effective GDNN-based framework for network optimization is still in its infancy. In the following we discuss what are, in our opinion, some of the most important challenges raised by this exciting scenario.

A. Data collection and sharing

A key enabler for the optimization approach proposed in this paper is the ability to collect data from various layers of the protocol stack, the environment, and even the final user, and to share these data at the network level. In general, we envision several types of data that can be collected, namely:

i) intra-device data (collected within each single device, e.g., protocol parameters or location), to be used in local optimizations (e.g., energy efficiency of a node);
ii) inter-device data (exchanged between devices, e.g., traffic patterns or queue lengths at routers), to be used for optimization on a wider scale (e.g., maximization of the number of flows the system can serve);
iii) user-profile data (which represent the users preferences) to define the Quality of Experience objective function to be used in the optimization.

Finding which data is most useful in the GDNN and for network optimization, studying the granularity and frequency at which these data need to be collected, and defining practical methods for representing, storing and retrieving such data at both the device and the system level are all open research problems that have to be solved.

B. Data representation and synchronization

Another open issue of the proposed optimization framework is the choice of the format of the data patterns that should be given as input to the GDNNs. In the context of network optimization, indeed, the sensory data might come in many different formats, which should nevertheless be encoded as activation values on the input layer of the network. This implies the need to carefully design a variety of encoding modules that should be used to transform the collected data into a unified representation, which should preserve as much as possible the inherent structure present in the data. This problem becomes even more challenging when considering data coming from heterogeneous devices and/or abstraction layers and from different time scales, or collected with different sampling frequencies or even asynchronously. Therefore, some effort shall be devoted to the identification of a solution for the data representation problem, which will also allow to better understand which are the most critical dimensions of the data domain (i.e., the most informative input signals) that should be given to the learning system.

C. Exploiting long-term spatio-temporal behaviors

An important component of cognition is the ability to adapt based on behaviors that have been observed and learned in the past and are likely to repeat again. How to include knowledge of the long-term spatio-temporal behavior of the network parameters (such as congestion or channel characteristics) into the optimization framework is an open research issue. Different strategies to include the time dimension into the generative models can be considered. A possible way is to build input vectors that collect the system parameters sampled at different time scales, in order to provide a representative example of the time evolution of the system. Another promising possibility is to use more complex generative models that are inherently sequential, such as the Recurrent Temporal Restricted Boltzmann Machine [38] or similar models that can be even combined into hierarchical architectures [16]. Further research is needed to gain a deeper understanding of these and other approaches and to find the best solution for the different optimization goals.

D. Multi-objective optimization strategies

The final objective of COBANETS shall be the automatic management of complex systems, in which individual agents may have both selfish objectives and common social goals to pursue (the latter possibly encouraged by game-theoretic or trust and reputation-based incentives). This problem may be approached using multi-objective optimization techniques, or by properly defining utility functions that jointly account for multiple objectives, appropriately weighed, or through a hierarchical organization of the goals. The specific properties of the generative models shall likely be combined with reinforcement learning mechanisms to automatically learn the best strategies in such a complex scenario.

E. Identification of domain-specific deep architectures

A crucial aspect to improve the performance and the scalability of many learning systems is to identify a useful set of constraints that can facilitate learning, for instance by reducing the complexity of the model or by improving convergence. For example, the most successful deep architectures for visual object recognition have been designed to exploit the strong local spatial correlation found in natural images [39]. It is therefore of interest to investigate how the distinguishing characteristics of telecommunication network signals can influence the deep architectures for learning-based

optimizations. For example, the deep network architectures may be designed to better process data with strong spatio-temporal correlation, or to account for the interdependencies among network elements induced by network topology. Moreover, when the deep network is fed with data originated by multiple devices interconnected through a communication network, there may be significant communication delays or even packet losses, thereby posing concrete challenges to a learning system that is usually expected to receive "clean" and reliable training patterns. These problems represent a less studied field of research that can potentially generate new insights and advances also in the machine learning domain.

F. Alternative building blocks for unsupervised learning

Generative models can take different forms, such as autoencoders, restricted Boltzmann machines and, more generally, energy-based models. Most of these models obtain similar performance in canonical machine learning experimental evaluations [40], but with different computational properties. Therefore, an interesting research subject will be the systematic study of the strengths and weaknesses of each approach in light of the considered optimization framework, and the investigation of which regularization techniques are more effective with the type of data and tasks required in such scenarios.

G. Knowledge distribution across network elements

A centralized management system may become the bottleneck of the optimization framework, in which case it would be preferable to distribute the optimization tasks to different network elements that should nevertheless be able to perform optimizations according to a global view of the networking environment. Some interesting recent studies have shown that the performance level obtained by very-large-scale deep neural networks in supervised classification tasks can be replicated in much smaller learning modules (*model compression*), such as simple networks with only one hidden layer, if we use as training labels the *soft-labels* produced as output by the large-scale deep network [42]. This intriguing result motivates further research about how to possibly create "lightweight" processing nodes that can support efficient optimization in a highly distributed system. Moreover, distributing the generative model over multiple nodes might be a valuable approach to speed up learning and inference tasks via efficient parallelization [41]. This feature is even more appealing considering that modern mobile devices (e.g., smartphones or notebooks) are equipped with powerful computing hardware, as discussed later on.

H. Security aspects

In our cognition-based approach, the network will need to continuously collect large amounts of data, apply a learning process to it, and take actions as a result, which will make the confidentiality of the original data, as well as that of the "reasoned" outcome, much more important than in traditional TCP/IP networks. For example, by changing behavior and observing how the network reacts, a user may obtain private information of others [43]. An open issue is to find the proper tradeoff between confidentiality and effectiveness of the proposed solutions, also considering possible de-anonymization techniques and privacy attacks based on machine learning [44]. Another problem is to design solutions to assess the trustworthiness of both peers and data (against attacks to either evade security checks or poison the learning process with fake data), as well as to make the learning process resilient to malicious attacks (e.g., based on Adversarial Machine Learning [45]). These are just a few examples of a number of innovative and challenging research problems concerning the security of cognitive system, which shall also include data confidentiality, trustworthiness, and resiliency to attacks.

I. Implementation and prototyping

Since the proposed approach gravitates around the possibility of turning the complexity of the system into an advantage, rather than an obstacle, by exploiting the inner capabilities of deep learning generative architectures to capture and discriminate hidden features of the complex multidimensional signals that are observed in real scenarios, the availability of large datasets of experimental data is essential for the proper design of a cognition-based network. While some testbeds capable of collecting system-wide cross-layer parameters have been proposed (e.g., [46]), the real-time testing of machine learning algorithms on experimental data has not yet been systematically addressed. Moreover, further research is required to exploit the new parallel computing framework for common graphic processor units (GPUs) on mobile devices to run complex machine learning algorithms in real-time and at affordable prices.

V. Conclusions

In this paper we advocated Generative Deep Neural Networks (GDNNs) as the key building block of a new generation of cognition-empowered networks and systems, where the ability of GDNNs to extract richer context representations will be combined with different kinds of machine learning techniques to realize specific tasks, and will be integrated with the Software Define Networking and Network Function Virtualization paradigms to enable a flexible actuation and management of complex systems.

Although some preliminary results are very encouraging, the potential of the proposed approach is still to be discovered, and a number of interesting interdisciplinary research issues need to be addressed. A possible way to tackle these exciting challenges is to approach the problem gradually, progressively widening the scope of the network optimization goal. The first fundamental step shall consist in gaining a deeper understanding of the potential of the generative deep learning approach to model and optimize specific network functionalities, such as resource allocation at the PHY layer, setting of MAC parameters, scheduling, routing, traffic source modeling, and so on. Supported by a solid theoretical and experimental foundation, it will then be possible to develop a generative deep learning approach to system-level optimization. This phase will require to design GDNNs capable of representing all the relevant functionalities that concur in determining the system performance, and to address the most critical and

challenging issues related to the scalability of the approach, the multi-objective optimization of the system parameters, the coordination of the different functionalities and network elements, and the implementation of the planned actions.

Besides leading to novel methods for the optimization of communication systems, this research may stimulate innovation in cognitive science and machine learning as well, leading to the development of new learning techniques that need to obey different constraints and boundary conditions than traditionally found in those areas. Therefore, we believe that the COBANETS concept may pave the way to new research avenues that intersect multiple sectors in cognitive science and information and communication engineering, with the potential of leading to disruptive innovation in these fields, with unpredictable effects on other fields that may benefit from the stimuli and the change of perspective brought about by the proposed vision.

REFERENCES

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, *Five disruptive technology directions for 5G.* IEEE Communications Magazine, 52(2), 74-80, 2015.

[2] W. H. Chin, Z. Fan, and R. Haines, *Emerging technologies and research challenges for 5G wireless networks.* IEEE Wireless Communications, 21(2), pp 106-112, 2014.

[3] P. Demestichas, et al. *5G on the horizon: key challenges for the radio-access network.*, IEEE Vehicular Technology Magazine, 8.3. pp 47-53, 2013.

[4] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, M. Zorzi, *The challenges of M2M massive access in wireless cellular networks*, Digital Communications and Networks, v. 1, n. 1, pp 1-19, 2015.

[5] S. Y. Lien, and K. C. Chen, *Massive access management for QoS guarantees in 3GPP machine-to-machine communications.* IEEE Communications Letters, 15(3), pp 311-313, 2011.

[6] F. Guidolin, I. Pappalardo, A. Zanella, and M. Zorzi, M. *A Markov-based framework for handover optimization in HetNets.* In the IEEE Ad Hoc Networking Workshop (MED-HOC-NET), pp. 134-139, 2014.

[7] D. Lopez-Perez, I. Güvenc, and X. Chu, *Mobility management challenges in 3GPP heterogeneous networks.* IEEE Communications Magazine, 60(12), pp 70-78, 2012.

[8] A. Khan, L. Sun, and E. Ifeachor, *QoE prediction model and its application in video quality adaptation over UMTS networks.* IEEE Transactions on Multimedia, 14(2), pp 431-442, 2012.

[9] S. Thakolsri,W. Kellerer, and E. Steinbach, *QoE-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation.* In IEEE International Conference on Communications (ICC), pp. 1-6, 2011.

[10] D. Kreutz, F. M. Ramos, P. Esteves Verissimo, C. Esteve Rothenberg, S. Azodolmolky, S., and S. Uhlig, *Software-defined networking: A comprehensive survey.* Proceedings of the IEEE, 103(1), 14-76. 2015.

[11] S. Haykin, *Cognitive radio: brain-empowered wireless communications,* IEEE Journal on Selected Areas in Communications, vol. 23, no. 2, pp. 201-220, Feb. 2005.

[12] J. Mitola III, *Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio,* Ph.D. dissertation, Royal Institute of Technology (KTH), Sweden, May 2000.

[13] N. Devroye, M. Vu, V. Tarokh, *Cognitive radio networks,* IEEE Signal Proc. Mag., Nov. 2008.

[14] F. H. P. Fitzek, M. Katz, *Cognitive Wireless Networks,* Springer, 2007.

[15] R. W. Thomas, D. H. Friend, L. A. DaSilva, A. B. MacKenzie, *Cognitive networks: Adaptation and learning to achieve end-to-end performance objectives,* IEEE Commun. Mag., pp. 51-57, Dec. 2006.

[16] G. Taylor, G. E. Hinton, *Factored conditional restricted Boltzmann Machines for modeling motion style,* Intern. Conference on Machine Learning, pp. 1025-1032, New York, NY, USA: ACM Press, 2009.

[17] D. Clark, C. Partridge, C. Ramming, J. Wroclawski, *A knowledge plane for the Internet,* ACM SIGCOMM, 2003

[18] M. Bkassiny, Y. Li, S.K. Jayaweera, *A Survey on Machine-Learning Techniques in Cognitive Radios,* IEEE Surveys and Tutorials, 2013.

[19] C. Clancy, J. Hecker, E. Stuntebeck, T. OShea, *Applications of machine learning to cognitive radio networks,* IEEE Wireless Communications, vol. 14, no. 4, pp. 47-52, 2007.

[20] A. Forster, *Machine learning techniques applied to wireless ad-hoc networks: Guide and survey,* IEEE 3rd Intern. Conf. on Intelligent Sensors, Sensor Networks and Inform. (ISSNIP), pp. 365-370, 2007.

[21] T. T. Nguyen, G. Armitage, *A survey of techniques for internet traffic classification using machine learning,* IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56-76, Oct. 2008.

[22] K. Winstein, H. Balakrishnan, *Tcp ex machina: Computer-generated congestion control,* Computer Communication Review (ACM SIGCOMM), vol. 43, no. 4, pp. 123-134, 2013.

[23] G. E. Hinton, T.J. Sejnowski, Unsupervised Learning: Foundations of Neural Computation. Cambridge, MA: MIT Press, 1999.

[24] Y. Bengio, *Deep learning of representations for unsupervised and transfer learning,* International Conference on Machine Learning, vol. 7, pp. 1-20, 2011.

[25] G. E. Hinton, Z. Gharahami, *Generative models for discovering sparse distributed representations,* Phil. Trans. of the Royal Society of London Series B: Biological Sciences, vol. 352, pp. 1177-1190, 1997.

[26] D. Koller, N. Friedman, *Probabilistic graphical models: principles and techniques,* MIT, 2009.

[27] G. E. Hinton, R. Salakhutdinov, *Reducing the dimensionality of data with neural networks,* Science, vol. 313, no. 5786, pp. 504-7, 2006.

[28] Y. Bengio, *Learning Deep Architectures,* AI Now Publishers Inc., 2009.

[29] Y. LeCun, Y. Bengio, G. Hinton, *Deep learning,* Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, et al., *Human-level control through deep reinforcement learning,* Nature, vol. 518, no. 7540, pp. 529-533, 2015.

[31] Y. Bengio, A. Courville, P. Vincent, *Representation Learning: A Review and New Perspectives,* IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1798-1828, 2013.

[32] M. Zorzi, A. Testolin, I. Stoianov, *Modeling language and cognition with deep unsupervised learning: a tutorial overview.* Frontiers in Psychology, vol. 4, 515, 2013.

[33] K. J. Friston, R. Adams, L. Perrinet, M. Breakspear, *Perceptions as hypotheses: saccades as experiments,* Frontiers in Psychology, vol. 3, 151, May 2012.

[34] D. Munaretto, D. Zucchetto, A. Zanella, and M. Zorzi, *Data-driven QoE optimization techniques for multi-user wireless networks.* In IEEE International Conference on Computing, Networking and Communications (ICNC), pp. 653-657, 2015.

[35] A. Testolin, M. Zanforlin, M. De Filippo De Grazia, D. Munaretto, A. Zanella, M. Zorzi, M. Zorzi, *A machine learning approach to QoE-based video admission control and resource allocation in wireless systems.* In the proceedings of IEEE Annual Mediterranean Ad Hoc Networking Workshop (MED-HOC-NET), pp. 31-38, 2014.

[36] M. Zanforlin, D. Munaretto, A. Zanella, M. Zorzi, *SSIM-based video admission control and resource allocation algorithms.* In the proceedings of International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pp. 656-661, 2014

[37] N. Srivastava and R.R. Salakhutdinov, *Multimodal learning with deep Boltzmann machines,* In Advances in neural information processing systems, pp. 2222-2230, 2012.

[38] I. Sutskever, G. E. Hinton, G. Taylor, *The recurrent temporal restricted Boltzmann machine,* Advances in Neural Information Processing Systems, vol. 20, pp. 1601-1608, 2008.

[39] A. Krizhevsky, I. Sutskever, G. E. Hinton, *ImageNet classification with deep convolutional neural networks,* Advances in Neural Information Processing Systems, vol. 24, pp. 609-616, 2012.

[40] Y. Bengio, O. Delalleau, *Justifying and generalizing contrastive divergence,* Neural Computation, vol. 14, no. 6, pp. 1601-1621, 2009.

[41] J. Dean, G. Corrado, R. Monga, et al., *Large Scale Distributed Deep Networks,* Advances in Neural Information Processing Systems, vol. 24, pp. 1-9, 2012.

[42] J. Ba, R. Caruana, *Do Deep Nets Really Need to be Deep?,* Advances in Neural Information Processing Systems, vol. 27, pp. 2654-2662, 2014.

[43] G. Ács, M. Conti, P. Gasti, C. Ghali, G. Tsudik, *Cache Privacy in Name-Data Networking,* 33rd Intern. Conf. on Distributed Comput. Systems (ICDCS), pp. 41-51, Jul. 8-11, 2013, Philadephia, PA, USA.

[44] M. Conti, L. V. Mancini, R. Spolaor, N. V. Verde, *Can't you hear me knocking: Identification of user actions on Android apps via traffic analysis,* 5th ACM SIGSAC CODASPY, Mar. 2-4, 2015, USA.

[45] M. Brad, A. Kantchelian, S. Afroz, R. Bachwani, E. Dauber, L. Huang, M. Carl Tschantz, A. D. Joseph, and J. D. Tygar. *Adversarial active learning.* In Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop, pp. 3-14. ACM, 2014.

[46] M. Danieletto, G. Quer, R.R. Rao, M. Zorzi, *CARMEN: a cognitive networking testbed on Android OS devices,* IEEE Comm. Mag., Sep. 2014