ICRI-CI 2017 Retreat – Agenda Day 1 – Abstracts (updated 30-Apr-2017)

Start	Durat.	End	Session	Speaker	Title/Chair
8:15	0:45	9:00	**Registration, Gathering, Breakfast**		
9:00	1:00	10:00	**Opening + Keynote**		
9:00	0:20	9:20		Ronny R+ Shalom G	**Opening Notes**
9:20	0:40	10:00		Keynote - Gadi Singer	**The ascent of AI, an Intel perspective** The talk will address the rapid changes in AI technologies and capabilities and the transformation they drive in computing. It will also describe Intel's key programs in Deep Learning and Cognitive Computing.
10:00	2:30	12:30	**Deep Learning I**		**Ronny Ronen**
10:00	0:30	10:30		Tali Tishby	**Opening the black box of Deep Neural Networks via Information: A deeper theory and some new algorithms** Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. In our previous work we proposed to analyze DNNs in the Information-Plane; i.e., the plane of the Mutual Information values that each layer preserves on the input and output variables. We suggested that the commonly used Stochastic Gradient Descent (SGD) methods for training DNN's effectively optimize the Information Bottleneck (IB) tradeoff between representation compression and label prediction, successively, for each layer. In this talk we first demonstrate the effectiveness of the Information-Plane visualization of DNNs. We then show that the Stochastic Gradient Descent (SGD) epochs exhibit two distinct phases: fast empirical error minimization (ERM, training label prediction) followed by slow representation compression, for each layer. During the ERM phase, the batch fluctuations of the estimated error gradients are small and the weights go through a fast drift to a small training error configuration. Most of the training epochs, however, are spent on the second phase, where the gradient fluctuations are much larger than their means (small gradient SNR) and the weights go through a slow diffusion process to highly compressed representations. We argue that this second phase, in which the gradients are very noisy, is the key to the success of Deep Learning, as they push the layers to (one of the many) optimal Information Bottleneck representations. This new important insight is consistent with other recent studies on the role of the noise in SGD optimization, and provide a new theoretical understanding of the computational benefits of the many hidden layers, their optimal organization and design principles. Based on joint works with Ravid Schwartz-Ziv and Noga Zaslavsky
10:30	0:30	11:00	Break + **Posters**		
11:00	0:30	11:30		Amnon Shashua	**Expressive efficiency and inductive bias of convolutional networks: the use of hierarchical tensor decompositions for network design and analysis** The driving force behind convolutional networks - the most successful deep learning architecture to date, is their expressive power. Despite its wide acceptance and vast empirical evidence, formal analyses supporting this belief are scarce. The primary notions for formally reasoning about expressiveness are efficiency and inductive bias. Efficiency refers to the ability of a network architecture to realize functions that require an alternative architecture to be much larger. Inductive bias refers to the prioritization of some functions over others given prior knowledge regarding a task at hand. Through an equivalence to hierarchical tensor decompositions, we study the expressive efficiency and inductive bias of various architectural features in convolutional networks (depth, width, pooling geometry, inter-connectivity, overlapping operations etc.). Our results shed light on the demonstrated effectiveness of convolutional networks, and in addition, provide new tools for network design.

11:30	0:30	12:00		Ohad Shamir	**Failures of Gradient-Based Deep Learning** In recent years, deep learning has become the go-to solution for a broad range of applications, with a long list of success stories. However, it is important, for both theoreticians and practitioners, to also understand the associated difficulties and limitations. In this work, we describe several simple problems for which commonly-used deep learning approaches either fail or suffer from significant difficulties. We illustrate these empirically, as well as provide theoretical insights explaining their source and (sometimes) how they can be remedied. Joint work with Shai Shalev-Shwartz and Shaked Shammah
12:00	0:30	12:30		Shie Mannor	**End-to-end Deep Imitation Learning** Generative adversarial learning is a popular approach to training deep learning architectures. The general idea is to maintain an oracle D that discriminates between the expert's data distribution and that of the generative model G. The generative model is trained to capture the expert's distribution by maximizing the probability of D misclassifying the data it generates. Overall, the system is differentiable end-to-end and is trained using back-propagation. This type of learning was successfully applied to the problem of policy imitation in a model-free setup. However, a model-free approach does not allow the system to be differentiable, which requires the use of high-variance gradient estimations. We devise model-based adversarial imitation learning for imitation learning. We show how to use a forward model to guarantee that the system is fully differentiable, which enables us to train policies using the (stochastic) gradient of D. Moreover, our approach requires relatively few interactions with the environment, and fewer hyper-parameters to tune.
12:30	2:00	14:30	Lunch + **Posters**		

14:30	3:00	17:30	**Deep Learning II + Visual**		**Ronny Ronen**
14:30	0:30	15:00		Lior Wolf	**Stereo Matching, Optical Flow, and Filling the Gaps** I will describe three projects around the topic of motions. (1) Improved stereo matching with constant highway networks and reflective confidence learning. (2) A brain inspired neural network for optical flow dense interpolation. (3) Training multiple strategies with one network. In addition, we will describe a state of the art system for matching images and text.
15:00	0:30	15:30		Shai Shalev Schwartz	**Deep reinforcement learning for driving policy** Technology for autonomous driving requires both "sensing" (understanding the environment) and "acting" (moving the car appropriately). The talk will focus on the "acting" part, which we call a "driving policy". Driving policy is a challenging task where the host vehicle must apply sophisticated negotiation skills with other road users when overtaking, giving way, merging, taking left and right turns and while pushing ahead in unstructured urban roadways. Moreover, one must balance between unexpected behavior of other drivers/pedestrians and at the same time not to be too defensive so that normal traffic flow is maintained. We discuss the application of reinforcement learning to driving policy, focusing on specific challenges to this domain, including learning with safety guarantees and reinforcement learning beyond MDPs.
15:30	0:30	16:00	Break + **Posters**		
16:00	0:30	16:30		Hayit Greenspan	**Deep Learning in Medical Imaging: The Data Challenge** In this talk I will introduce several applications of deep learning methods in medical imaging. The main challenge in medical applications is the limited availability of expert-labeled data. I will discuss the issues involved and several schemes that have evolved to overcome them and enable the use of networks in the medical domain. Specific applications that I will present include Brain MRI lesion detection, CT liver segmentation and lesion detection as well as Chest X-ray pathology categorization.
16:30	0:30	17:00		Daphna Weinshall	**Implicit Media Tagging and Affect Prediction from RGB-D video of spontaneous facial expressions** I will discuss a method that automatically evaluates emotional response from spontaneous facial activity. The method is based on the inferred activity of facial muscles over time, as automatically obtained from an RGB-D video recording of spontaneous facial activity. The contribution of this work is two-fold: First, we constructed a database of publicly available short video clips, which elicit a strong emotional response in a consistent manner across different individuals. Each video was tagged by its characteristic emotional response along 4 scales: Valence, Arousal, Likability and Rewatch (the desire to watch again). The second contribution is a two-step prediction method, based on learning, which was trained and tested using this database of tagged video clips. The proposed method was able to successfully predict the aforementioned 4 dimensional representation of affect, achieving high correlation (0.87-0.95) between the predicted scores and the affect tags. As part of the prediction algorithm we identified the period of strongest emotional response in the viewing recordings, in a method that was blind to the video clip being watched, showing high agreement between independent viewers. Finally, inspection of the relative contribution of different feature types to the prediction process revealed that temporal facets contributed more to the prediction of individual affect than to media tags.
17:00	0:30	17:30		Yair Weiss	**The Return of the Gating Network: Combining Discriminative and Generative Training in models of RGBD Images** In recent years, approaches based on machine learning have achieved state-of-the-art performance on image restoration problems. Successful approaches include both generative models of natural images as well as discriminative training of deep neural networks. Discriminative training of feed forward architectures allows explicit control over the computational cost of performing restoration and therefore often leads to better performance at the same cost at run time. In contrast, generative models have the advantage that they can be trained once and then adapted to any image restoration task by a simple use of Bayes' rule. In this paper we show how to combine the strengths of both approaches by training a discriminative, feed-forward architecture to predict the state of latent variables in a generative model of RGBD image patches. We apply this idea to the very successful Gaussian Mixture Model (GMM). We show that it is possible to achieve comparable performance as the original GMM model but with two orders of magnitude improvement in run time while maintaining the advantage of generative models.
17:30	1:30	19:00	**Reception + Posters**		

ICRI-CI 2017 Retreat – Agenda Day 2 – Abstracts

Start	Durat.	End	Session	Speaker	Title/Chair
8:15	0:45	9:00	**Registration, Gathering, Breakfast**		
9:00	3:00	12:00	**Architecture**		**Debbie Marr**
9:00	0:30	9:30		Debbie Marr	**Architecture**
9:30	0:30	10:00		Uri weiser	**Effective usage of system's resources – when/where should we use In-Place-Processing** The era of Big Data processing is already here. Many Big Data applications' phases exhibit non-temporal locality memory accesses. However, modern computing systems perform well under workloads that exhibit temporal data locality accesses. Industry and academia are already exploiting the potential of In-Place-Processing. Our research aim at identification of programs' phases and systems' characteristics that should lead towards the effective usage of In-Place-Processing. In this talk we will present the progress of our research since last year's ICRI-CI retreat talk. We will show results of real system's evaluations that support our analytical model, based on real applications' computational phases. Furthermore, the next steps of this research will be presented.
10:00	0:30	10:30		Ran Ginosar	**PRinS: Processing-in-Storage Using Resistive CAM** Near-data in-storage processing research has been gaining momentum in recent years. Typical processing-in-storage architecture places a single or several processing cores inside the storage and allows data processing without transferring it to the host CPU. Since this approach replicates von-Neumann architecture inside storage, it is exposed to the problems faced by von-Neumann architectures, especially the bandwidth wall. We present a novel processing-in-storage system based on Resistive Content Addressable Memory (RCAM). RCAM functions simultaneously as a storage and a massively parallel associative processor. RCAM processing-in-storage resolves the bandwidth wall faced by conventional processing-in-storage architectures by keeping the computing inside the storage arrays, thus implementing in-data, rather than near-data, processing. We show that RCAM based processing-in-storage architecture may outperform existing in-storage designs and accelerator based designs. RCAM processing-in-storage implementation of k-Means achieves speedup of 4.6—68 relative to CPU, GPU and FPGA based solutions. For k-Nearest Neighbors, RCAM processing-in-storage achieves speedup of 17.9—17,470 and for Smith-Waterman sequence alignment it reaches speedup of almost 5 over a GPU cluster based solution.
10:30	0:30	11:00	Break + **Posters**		
11:00	0:30	11:30		Shahar Kvatinsky	**mMPU: Memristive Memory Processing Unit** Memristive technologies are attractive candidates to replace conventional memory technologies, and can also be used to perform logic and arithmetic operations using a technique called 'stateful logic.' Combining data storage and computation in the memory array enables a novel non-von Neumann architecture, where both the operations are performed within a memristive Memory Processing Unit (mMPU). mMPU relies on adding computing capabilities to the memristive memory cells without changing the basic memory array structure. The use of an mMPU alleviates the primary restriction on performance and energy in von Neumann machine, which is the data transfer between CPU and memory. This talk focuses on the various aspects of mMPU. We will discuss its architecture and implications on the computing system and software, as well as examining the micro-architectural aspects. We will show how to design the mMPU controller and how different sequence of computing operations in an mMPU can be automatically optimized as sequences of basic Memristor Aided Logic (MAGIC) NOR and NOT operations. We will present examples of applications that can benefit from processing within memristive memory and show how adding mMPU to conventional computing systems substantially improves the system performance and energy efficiency.
11:30	0:30	12:00		Yoav Etsion	**Memory Prefetching with Neural-Networks: Challenges and Insights** Our research examines the use of neural networks (NNs) for data cache prefetching. We categorize state-of-the-art prefetchers based on the access patterns they target and show that a single NN can learn the distinct patterns, albeit with different convergence times. We demonstrate that, conceptually, a NN prefetcher can outperform a collection of prefetchers. We also highlight some implementation challenges in turning the conceptual NN prefetcher into a reality.
12:00	1:30	13:30	Lunch + **Posters**		

13:30	3:00	16:30	**Conversational Understanding**		**Moshe Wasserblat**
13:30	0:30	14:00		Moshe Wasserblat	**Conversational Understanding**
14:00	0:30	14:30		Ronen Feldman	**High Performance Information Extraction - Techniques and Applications** VIP (Visual Information Extraction Platform) is a development and evaluation tool for creating linguistic models by utilizing supervised and unsupervised machine learning methods and visual pattern definitions. The text mining engine that utilizes the linguistic models processes and analyzes any type of documents and translates it into structured data sets of relations between entities and sentiments tied to events. VIP uses state-of-the-art Natural Language Processing (NLP) and machine learning techniques to process and analyze text. VIP has several unique features: first, its visual properties allow users to easily interact with the system, without requiring them to have programming skills or write code. Through this unique interface, users can view the extracted sentiments, relations and events, and determine whether they are correct. Users can, furthermore, annotate missing events, relations and sentiments, and add new event types and their associated extraction rules. Second, VIP uses a proprietary leading edge pattern language, designed to identify relevant linguistic patterns in context-free grammars involving sentiments, relations or events. VIP is utilizing a pipeline architecture of several NLP and ML components that enable the user to mix and match between the various components. It is easy to add to the architecture other open source components such as parsers, NER (Named Entity Recognition), POS (Part of Speech) or Word2Vec modules for finding domain specific synonyms for terms and words.
14:30	0:30	15:00		Yoav Goldberg	**Accurate parsing and beyond** Syntactic parsing is an essential component in language understanding. Using neural networks and representation-learning techniques, we manage to substantially improve dependency parsing accuracies. I will describe our work on the BIST parser, that does this. However, accurate dependency parsing is not enough, as many linguistic phenomena of interest are not expressed in the resulting syntactic structure. I will show some of these cases, which my group is recently starting to tackle.
15:00	0:30	15:30	Break + **Posters**		
15:30	0:30	16:00		Ido Dagan	**Open Knowledge Representation and Lexical Inference** In the first part of this talk, I will outline our broad research direction, termed Open Knowledge Representation (OKR). It aims to represent textual information from multiple texts in a consolidated manner, based on the available natural language vocabulary, without relying on pre-specified schemata or ontology. Our proposed structure merges co-referring individual proposition extractions, created in an Open-IE flavor, into a representation of consolidated entities and propositions, inspired by traditional knowledge graphs. In the second part of the talk, I will describe our recent advances in modeling lexical inferences, which are needed for the knowledge graph construction as well as for many other applications, including targeted applications in the context of the ICRI-CI project.
16:00	0:30	16:30		Roi Reichart	**Morphological and Multi-lingual Specialization of Vector Spaces and its Applications to Conversation Understanding** Vector space models for word and sentence embeddings have made a substantial impact on Natural Language Processing (NLP) over the last few years. However, for resource-poor languages and particularly for those that are morphologically rich, the quality of these models is substantially lower compared to languages like English. In this talk we describe a new algorithm, named ATTRACT-REPEL, for improving the semantic quality of word vectors by injecting constraints extracted from both manually constructed lexical resources and from simple hand-crafted rules. We show how to employ this model with monolingual and cross-lingual constraints from lexicons as well as with very simple manually crafted morphological rules. We demonstrate that the vector sets generated by a large variety of state-of-the-art word embedding models can be substantially improved, with the most noticeable gains observed for vector spaces in resource poor and morphologically rich languages. The quality of our vectors are demonstrated in both intrinsic and task based evaluation: we demonstrate how to incorporate them in a dialog state tracking system in a way that form a new state-of-the-art on that task.
16:30	0:15	16:45	**Wrap-up**		