Recent Advances at the Brain-Driven Computer Vision Workshop 2018

Simone Palazzo[1], Isaak Kavasidis[1], Dimitris Kastaniotis[2], and Stavros Dimitriadis[3]

[1] University of Catania, Italy
[2] University of Patras, Greece
[3] Cardiff University, United Kingdom
palazzosim@dieei.unict.it, kavasidis@dieei.unict.it, dkastaniotis@upatras.gr, DimitriadisS@cardiff.ac.uk

Abstract. The 1st edition of the Brain-Driven Computer Vision Workshop, held in Munich in conjunction with the European Conference on Computer Vision 2018, aimed at attracting, promoting and inspiring research on paradigms, methods and tools for computer vision driven or inspired by the human brain. While successful, in terms of the quality of received submissions and audience present at the event, the workshop emphasized some of the factors that currently limit research in this field. In this report, we discuss the success points of the workshop, the characteristics of the presented works, and our considerations on the state of current research and future directions of research in this topic.

Keywords: Brain-Driven Computer Vision, Biologically-Inspired Machine Learning

1 Introduction

In recent years, we have witnessed unprecedented advancements in the automatic analysis of visual data by computer algorithms, thanks to a series of factors which unleashed the potential of convolutional neural networks. Part of the reason of their success could be explained by the biologically-inspired design of such models: indeed, while the classic artificial neuron may only by an extreme simplification of the biological neuron, the increasing representational complexity learned by CNNs may be more faithful to the layered structure of the lower areas of the human visual cortex [11, 12]. However, the road towards achieving a degree of artificial emulation of the human visual system high enough to interpret an environment as humans do is still long: while we are able to identify low- to high-level visual patterns from images and videos, artificial models largely miss the human capability to make sense of this information, recognize semantic patterns, correlate to memory and experience, and so on. Additionally, even the kind of neural computational models that we employ are only loosely based of biological structures and connections: for example, human visual analysis transmits information across cortical brain regions in both feedforward and feedback patterns, with the latter basically missing from current artificial approaches [13–16]. Although some recent works [17] attempted to encode hierarchical predictions [18, 19] featuring a combination of feedforward, feedback and recurrent connections, while others explored methods to decode brain visual representations [20–24], we still lack a sufficient understanding of the underlying cognitive processes to emulate and transfer them to physiologically-motivated implementations.

The proposed workshop is inspired by the realization that understanding and speculating on the yet mostly unknown mechanisms used by the brain to process visual information and knowledge may be the key to further advance computer vision beyond the black-box paradigm of training from data in the hope to uncover the very same processes. The daunting yet fascinating challenge presented by this task calls for a largely multidisciplinary effort by research communities in the fields of artificial intelligence, machine learning, cognitive neuroscience, psychology, among others. The aim of the workshop was therefore to excite the study and development of paradigms, methods and tools for computer vision driven or inspired by the human brain, both as a computational model and a source of data, and to promote the diffusion of new benchmarks and evaluation protocols to support the scientific community in the pursuing of a better understanding of the brain processes underlying human visual perception and comprehension.

2 Taxonomy of Workshop Papers

One of the objectives of the Brain-Driven Computer Vision Workshop was to blend research in computer vision with neuroscience, and from this point of view the achieved results were only partly satisfactorily, as many works had a distinct computer-vision–oriented slant, with the "brain-driven" aspect mostly consisting in an algorithmic inspiration.

Hence, our overview of the works presented at the workshop separates them into two categories, based on the perspective of the study regarding the representation of visual information. Papers focusing on representation in biological neurons and on how these neurons respond to different visual stimuli will be discussed in Section 2.1; papers focusing on the design of biologically-inspired algorithms for processing digital information will be presented in 2.2.

2.1 Brain representation and human perception

Visual information processing in human and animal brains is one of the most interesting topics of neuroscience. Traditionally, these studies have revealed significant findings, which in many cases have helped in the development of state of the Computer Vision models [25], [26]. In brief, research has been focused on the encoding mechanisms inside visual cortex as well as on the two stream processing hypothesis (hierarchy of processing layers) — namely the ventral (shape processing) and the dorsal (motion processing) systems.

The similarities and differences between biological and artificial neurons is a long-time research topic in computational neuroscience. In [1], the authors discuss the ability of a neural network trained on a particular task to describe the behaviour of neurons that are dedicated to operate on the same task, with a focus on scene-parsing models, that are shown to better explain task-specific brain responses than scene classification models.

In [2], authors focus on the shape information processing (ventral stream) by analyzing the visual stimuli responses in V1 cortex of anesthetized naive Long-Evans rats. To this end, they recorded extracellular potentials and focused on low-level features (position of center of mass and luminosity). Extracellular potentials were initially filtered with a band-pass filter (0.5-11 KHz) followed by an Expectation-Maximization clustering algorithm to differentiate between spikes produced by different neurons. They analyzed these recordings using a number of clustering and classification techniques and concluded that both luminosity and position as well as the combination of the two, are naturally mapped in the V1.

In the spirit of [11], the authors of [3] target the problem of decoding visual representation from fMRI, by analyzing feature correlation through a regressive model trained from brain data and comparing several machine learning methods and similarity measures in order to maximize decoding accuracy. They found that a Multilayer Neural Network is able to best represent the non-linear relationship between a Deep CNN and the features of fMRI. Also, features from the whole Visual Cortex surpass the performance of individual cortices. Also, they observed that higher visual cortex areas surpass the lower visual cortex but also, in the lower visual cortex, V4 area surpasses all previous areas— which is quite reasonable as it is an area of the ventral system.

2.2 Brain-inspired representation learning

Researchers in computer vision have traditionally tried to mimic the behavior of the human brain, as it is the most obvious reference model for understanding the visual world.

In this workshop, we had two papers focusing on Capsule Networks [27]. Capsule Networks are inspired by the working mechanism of optic neurons of the human visual system, achieving a significant improvement regarding the ability to efficiently detect presence of a an object in a scene. In [4], the authors presented an architecture for generative adversarial networks (GANs) where the discriminator was modeled as a capsule network and the loss function was adapted to include the standard capsule margin loss. Results evaluated using a Generative Adversarial Metric and a Semi-supervised Classification showed an improvement as compared to GANs working with a regular CNN Discriminator network.

Capsule Nets do not explicitly model the relationship of output vector activations. In this manner in [5], the authors proposed three improvements on the standard capsule network architecture. They proposed a novel routing weight initialization technique, the exploitation of semantic relationships between the primary capsule activations using a densely connected conditional random field

(CRF), and a Cholesky transformation–based correlation module to learn a general priority scheme. These modifications gave promising results on the problem of multi-label classification. For the evaluation they incorporated the ADE20K dataset [28] , which has 200.000 images from 150 scenes and used the mean Average Precision Metric (mAP) to compare the regular Capsule Networks with the proposed scheme.

Push-pull inhibition is a phenomenon observed in neurons in the V1 area of the visual cortex, which suppresses the response of certain simple cells for stimuli of preferred orientation but of non-preferred contrast, improving the quality of delineation especially in images with spurious texture. The authors of [6] presented a delineation operator that implemented a pull-push inhibition mechanism improving robustness to noise in terms of spurious texture.

In [7], the authors tested the robustness of face recognition models to false positive and, in particular, to simulacra and pareidolia, two categories of psychological phenomena that allow humans to recognize particular objects (such as an arrangement of three points resembling two eyes and a mouth) as faces, and that can be interpreted as false positives triggered by human psychological peculiarities. Their results showed that state-of-the-art models were not robust against these particular types of false positives, confirming the gap between algorithmic and human-level performance.

One of the abilities of animals is to efficiently subitize, i.e., counting the number of objects in a scene. In [8], the authors discussed the intrinsic abilities of deep convolutional neural networks to perform this task. They showed that variational autoencoders were able to spontaneously perform subitizing after training without supervision on a large amount of images. Also, they studied the effect of the size of the object and they concluded that the learned representations are likely invariant to object area. This observation is aligned with recent studies on biological neural networks in cognitive neuroscience.

While it seems trivial for a human subject to identify different temporal segments in a video (i.e., grouping images that belong to the same video scene), it is not so when automated methods are employed. Also, neuroscience findings indicate that stimuli belonging to the same temporal context, are grouped together in clusters of communities formed inside a representational space [29]. Inspired by these findings, in [9], the authors showed a method for learning a representation suitable for the task of temporal video segmentation by using directed graphs that represent how the feature vectors of the images in a video are connected temporally. These temporal edges represent the temporal similarity of the images in the video and then mapped to an automatically learned feature space by employing both LSTM and 'vanilla' neural networks.

In [10], the authors adopted a biological-like pyramidal structure of neuron interconnections to create a model that was able to understand human emotions from sequences of pictures. Given that human facial expressions must be taken in context, the authors opted to consider image sequences instead of single pictures. The image sequences provided the model with more information indicating the actual emotion that the subject displayed. The method was tested on two different datasets obtaining very good results and demonstrating that temporal cues in the expression of emotion play a significant role.

3 Discussion

From an organization point of view, this first edition of the workshop was successful, in terms of both received submissions and, equally important, the number of people attending the workshop. It should be noted that most of the attendees were present at the keynote speech and at the poster session, while fewer participated in the final discussion at the end of the workshop.

As for the quality and content of the accepted submissions, while all were on topic (as defined by the list published in the call for papers[4]), Section 2 highlights a certain unbalance between the number of papers tackling the analysis of biological data in the attempt to gain a better understanding of the underlying processes, and those proposing purely-algorithmic approaches with an (sometimes loose) inspiration to neurocognitive mechanisms, with the latter being significantly more numerous. Of course, this might have been expected: working on brain data requires, first and foremost, the availability or accessibility of such data, which is something that not all research groups have, as specific and expensive technology (e.g., EEG and fMRI) are necessary. However, while brain-inspired approaches can certainly be useful, both as practical tools/applications and as a source of architectural ideas for artificial model, we believe that brain activity analysis and the consequent efforts to uncover how visual information is represented and processed in the brain will be the key factor to devising artificial models that fulfill the learning, generalization and adaptation gaps to human performance.

From this point of view, we were hoping to see submissions providing new brain activity datasets to the research community. Given the above-mentioned difficulty in neuroimaging data collection, dataset availability seems to be the main limiting factor in the application of modern machine learning techniques to brain activity understanding, prediction and emulation.

In view of these considerations, the success of this first edition of the workshop will certainly encourage us to continue the series with a new edition; however, we believe that a stricter topic policy will be enforced, trying to attract more submissions that work with and on neuroimaging datasets, in the attempt to push the boundaries of current research in brain visual representation learning, decoding and understanding.

List of Workshop papers

1. Dwivedi, K., Roig, G.: Navigational affordance cortical responses explained by scene-parsing model
2. Vascon, S., Parin, Y., Annavini, E., DAndola, M., Zoccolan, D., Pelillo, M.: Characterization of visual object representations in rat primary visual cortex

[4] http://www.upcv.upatras.gr/BDCV/CFP.html

3. Papadimitriou, A., Passalis, N., Tefas, A.: Decoding generic visual representations from human brain activity using machine learning
4. Jaiswal, A., AbdAlmageed, W., Wu, Y., Natarajan, P.: Capsulegan: Generative adversarial capsule network
5. Ramasinghe, S., Athuraliya, C., Khan, S.: A context-aware capsule network for multi-label classification
6. Strisciuglio, N., Azzopardi, G., Petkov, N.: Brain-inspired robust delineation operator
7. Natsume, R., Inoue, K., Fukuhara, Y., Yamamoto, S., Morishima, S., Kataoka, H.: Understanding fake faces
8. Wever, R., Runia, T.F.: Subitizing with variational autoencoders
9. Dias, C., Dimiccoli, M.: Learning event representations by encoding the temporal context
10. Nardo, E.D., Petrosino, A., Ullah, I.: Emop3d: A brain like pyramidal deep neural network for emotion recognition

References

11. Horikawa, T., Kamitani, Y.: Generic decoding of seen and imagined objects using hierarchical visual features. Nat Commun **8** (May 2017) 15037
12. Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A.: Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep **6** (06 2016) 27755
13. Bullier, J.: Integrated model of visual processing. Brain Res. Brain Res. Rev. **36**(2-3) (Oct 2001) 96–107
14. Kourtzi, Z., Connor, C.E.: Neural representations for object perception: structure, category, and adaptive coding. Annu. Rev. Neurosci. **34** (2011) 45–67
15. Kravitz, D.J., Saleem, K.S., Baker, C.I., Mishkin, M.: A new neural framework for visuospatial processing. Nat. Rev. Neurosci. **12**(4) (Apr 2011) 217–230
16. DiCarlo, J.J., Zoccolan, D., Rust, N.C.: How does the brain solve visual object recognition? Neuron **73**(3) (Feb 2012) 415–434
17. Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., Liu, Z.: Deep predictive coding network for object recognition. In Dy, J., Krause, A., eds.: Proceedings of the 35th International Conference on Machine Learning. Volume 80 of Proceedings of Machine Learning Research., Stockholmsmssan, Stockholm Sweden, PMLR (10–15 Jul 2018) 5266–5275
18. Clark, A.: Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci **36**(3) (Jun 2013) 181–204
19. Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J.: Canonical microcircuits for predictive coding. Neuron **76**(4) (Nov 2012) 695–711
20. Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Souly, N., Shah, M.: Deep Learning Human Mind for Automated Visual Classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (jul 2017) 4503–4511
21. Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Shah, M.: Generative adversarial networks conditioned by brain signals. In: 2017 IEEE International Conference on Computer Vision (ICCV). (Oct 2017) 3430–3438
22. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L.: Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. **21**(19) (Oct 2011) 1641–1646

23. Stansbury, D.E., Naselaris, T., Gallant, J.L.: Natural scene statistics account for the representation of scene categories in human visual cortex. Neuron **79**(5) (Sep 2013) 1025–1034
24. Kavasidis, I., Palazzo, S., Spampinato, C., Giordano, D., Shah, M.: Brain2image: Converting brain signals into images. In: Proceedings of the 2017 ACM on Multimedia Conference, ACM (2017) 1809–1817
25. B A Olshausen, D.J.F.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images (1997) Nature.
26. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, Ieee (2007) 1–8
27. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems 30, Curran Associates, Inc. (2017) 3856–3866
28. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017)
29. Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., Botvinick, M.M.: Neural representations of events arise from temporal community structure. Nature neuroscience **16**(4) (2013) 486