The Deep Learning Epilepsy Detection Challenge: design, implementation, and test of a new crowd-sourced AI challenge ecosystem

Isabell Kiral*, Subhrajit Roy*, Todd Mummert*, **Alan Braz***, Jason Tsay, Jianbin Tang, Umar Asif, Thomas Schaffter, Eren Mehmet, The IBM Epilepsy Consortium◊, Joseph Picone, Iyad Obeid, Bruno De Assis Marques, Stefan Maetschke, Rania Khalaf†, Michal Rosen-Zvi†, Gustavo Stolovitzky†, Mahtab Mirmomeni†, Stefan Harrer†

* These authors contributed equally to this work

† Corresponding authors: rkhalaf@us.ibm.com, rosen@il.ibm.com, gustavo@us.ibm.com, mahtabm@au1.ibm.com, sharrer@au.ibm.com

◊ Members of the IBM Epilepsy Consortium are listed in the Acknowledgements section

J. Picone and I. Obeid are with Temple University, USA. T. Schaffter is with Sage Bionetworks, USA. E. Mehmet is with the University of Illinois at Urbana-Champaign, USA. All other authors are with IBM Research in USA, Israel and Australia.

Introduction

This decade has seen an ever-growing number of scientific fields benefitting from the advances in machine learning technology and tooling. More recently, this trend reached the medical domain, with applications reaching from cancer diagnosis [1] to the development of brain-machine-interfaces [2]. While Kaggle has pioneered the crowd-sourcing of machine learning challenges to incentivise data scientists from around the world to advance algorithm and model design, the increasing complexity of problem statements demands of participants to be expert data scientists, deeply knowledgeable in at least one other scientific domain, and competent software engineers with access to large compute resources. People who match this description are few and far between, unfortunately leading to a shrinking pool of possible participants and a loss of experts dedicating their time to solving important problems. Participation is even further restricted in the context of any challenge run on confidential use cases or with sensitive data. Recently, we designed and ran a deep learning challenge to crowd-source the development of an automated labelling system for brain recordings, aiming to advance epilepsy research. A focus of this challenge, run internally in IBM, was the development of a platform that lowers the barrier of entry and therefore mitigates the risk of excluding interested parties from participating.

The challenge: enabling wide participation

With the goal to run a challenge that mobilises the largest possible pool of participants from IBM (global), we designed a use case around previous work in epileptic seizure prediction [3]. In this "Deep Learning Epilepsy Detection Challenge", participants were asked to develop an automatic labelling system to reduce the time a clinician would need to diagnose patients with epilepsy. Labelled training and blind validation data for the challenge were generously provided by Temple University Hospital (TUH) [4]. TUH also devised a novel scoring metric for the detection of seizures that was used as basis for algorithm evaluation [5].

In order to provide an experience with a low barrier of entry, we designed a generalisable challenge platform under the following principles:

1. No participant should need to have in-depth knowledge of the specific domain. (i.e. no participant should need to be a neuroscientist or epileptologist.)
2. No participant should need to be an expert data scientist.
3. No participant should need more than basic programming knowledge. (i.e. no participant should need to learn how to process fringe data formats and stream data efficiently.)
4. No participant should need to provide their own computing resources.

In addition to the above, our platform should further

- guide participants through the entire process from sign-up to model submission,
- facilitate collaboration, and
- provide instant feedback to the participants through data visualisation and intermediate online leaderboards.

The platform

The architecture of the platform that was designed and developed is shown in Figure 1. The entire system consists of a number of interacting components. **(1) A web portal** serves as the entry point to challenge participation, providing challenge information, such as timelines and challenge rules, and scientific background. The portal also facilitated the formation of teams and provided participants with an intermediate leaderboard of submitted results and a final leaderboard at the end of the challenge. **(2) IBM Watson Studio** [6] is the umbrella term for a number of services offered by IBM. Upon creation of a user account through the web portal, an IBM Watson Studio account was automatically created for each participant that allowed users access to IBM's Data Science Experience (DSX), the analytics engine Watson Machine Learning (WML), and IBM's Cloud Object Storage (COS) [7], all of which will be described in more detail in further sections. **(3) The user interface and starter kit** were hosted on IBM's Data Science Experience platform (DSX) and formed the main component for designing and testing models during the challenge. DSX allows for real-time collaboration on shared notebooks between team members. A starter kit in the form of a Python notebook, supporting the popular deep learning libraries TensorFLow [8] and PyTorch [9], was provided to all teams to guide them through the challenge process. Upon instantiation, the starter kit loaded necessary python libraries and custom functions for the invisible integration with COS and WML. In dedicated spots in the notebook, participants could write custom pre-processing code, machine learning models, and post-processing algorithms. The starter kit provided instant feedback about participants' custom routines through data visualisations. Using the notebook only, teams were able to run the code on WML, making use of a compute cluster of IBM's resources. The starter kit also enabled submission of the final code to a data storage to which only the challenge team had access. **(4) Watson Machine Learning** provided access to shared compute resources (GPUs). Code was bundled up automatically in the starter kit and deployed to and run on WML. WML in turn had access to shared storage from which it requested recorded data and to which it stored the participant's code and trained models. **(5) IBM's Cloud Object Storage** held the data for this challenge. Using the starter kit, participants could investigate their results as well as data samples in order to better design custom algorithms. **(6) Utility Functions** were loaded into the starter kit at instantiation. This set of functions included code to pre-process data into a more common format, to optimise streaming through the use of the NutsFlow and NutsML libraries [10], and to provide seamless access to the all IBM services used. Not captured in the diagram is the **final code evaluation**, which was conducted in an automated way as soon as code was submitted though the starter kit, minimising the burden on the challenge organising team.

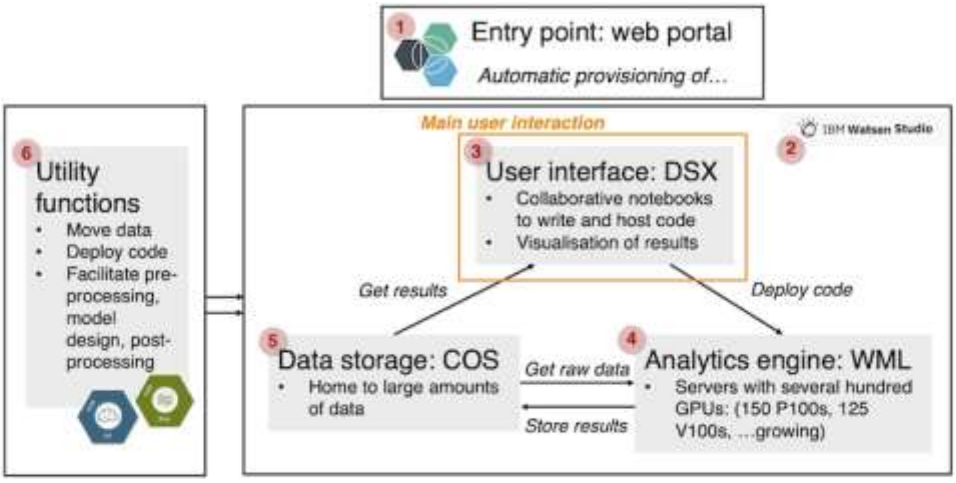

Figure 1: *High-level architecture of the challenge platform*

Measuring success

The competitive phase of the "Deep Learning Epilepsy Detection Challenge" ran for 6 months. Twenty-five teams, with a total number of 87 scientists and software engineers from 14 global locations participated. All participants made use of the starter kit we provided and ran algorithms on IBM's infrastructure WML. Seven teams persisted until the end of the challenge and submitted final solutions. The best performing solutions reached seizure detection performances which allow to reduce hundred-fold the time eliptologists need to annotate continuous EEG recordings. Thus, we expect the developed algorithms to aid in the diagnosis of epilepsy by significantly shortening manual labelling time. Detailed results are currently in preparation for publication.

Equally important to solving the scientific challenge, however, was to understand whether we managed to encourage participation from non-expert data scientists.

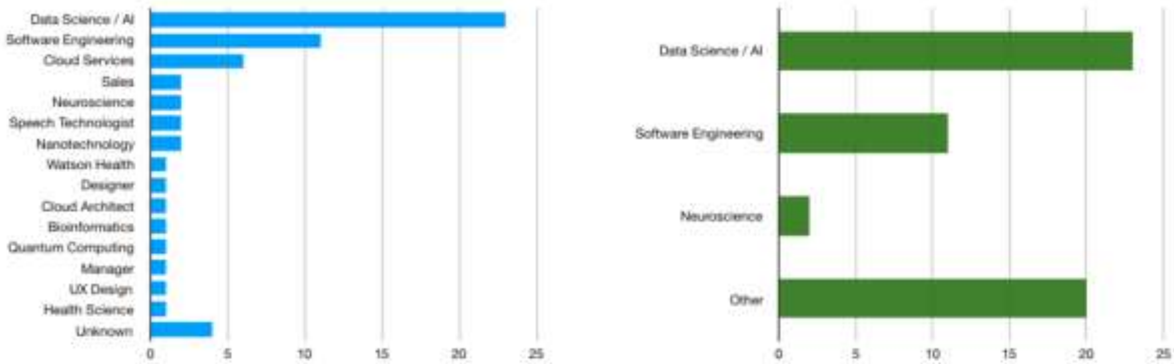

Figure 2: *Primary occupation as reported by challenge participants*

Out of the 40 participants for whom we have occupational information, 23 reported Data Science or AI as their main job description, 11 reported being a Software Engineer, and 2 people had expertise in Neuroscience. Figure 2 shows that participants had a variety of specialisations, including some that are in no way related to data science, software engineering, or neuroscience. No participant had deep knowledge and experience in data science, software engineering *and* neuroscience.

Conclusion

Given the growing complexity of data science problems and increasing dataset sizes, in order to solve these problems, it is imperative to enable collaboration between people with differences in expertise with a focus on inclusiveness and having a low barrier of entry. We designed, implemented, and tested a challenge platform to address exactly this. Using our platform, we ran a deep-learning challenge for epileptic seizure detection. 87 IBM employees from several business units including but not limited to IBM Research with a variety of skills, including sales and design, participated in this highly technical challenge.

References

[1] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), p.115.

[2] Nurse, E., Mashford, B.S., Yepes, A.J., Kiral-Kornek, I., Harrer, S. and Freestone, D.R., 2016, May. Decoding EEG and LFP signals using deep learning: heading TrueNorth. In *Proceedings of the ACM International Conference on Computing Frontiers* (pp. 259-266). ACM.

[3] Kiral-Kornek, I., Roy, S., Nurse, E., Mashford, B., Karoly, P., Carroll, T., Payne, D., Saha, S., Baldassano, S., O'Brien, T., Grayden, D., Freestone, D., and Harrer, S., 2018. Epileptic seizure prediction using big data and deep learning: toward a mobile system. EBioMedicine, 27, pp.103-111.

[4] Obeid, I. and Picone, J., 2016. The Temple University Hospital EEG data corpus. *Frontiers in neuroscience*, *10*, p.196.

[5] Ziyabari, S., Shah, V., Golmohammadi, M., Obeid, I. and Picone, J., 2017. Objective evaluation metrics for automatic classification of EEG events. *arXiv preprint arXiv:1712.10107*.

[6] https://cloud.ibm.com/catalog/services/watson-studio

[7] https://www.ibm.com/cloud/object-storage

[8] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (pp. 265-283).

[9] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A., 2017. Automatic differentiation in pytorch.

[10] Maetschke, S., Tennakoon, R., Vecchiola, C. and Garnavi, R., 2017. nuts-flow/ml: data pre-processing for deep learning. *arXiv preprint arXiv:1708.06046*.

Acknowledgements

The authors would like to thank all challenge participants as well as testers of the challenge platform.

The IBM Epilepsy Consortium consists of the following members: Hiroki Yanagisawa, Toshiya Iwamori, Piyush Madan, Yong Qin, Li Ma, Wei Lian Ti, Wen Liu, Jing Mei, Sharon Hensley, Rachita Chandra, Paul Hake, Richard Hennessy, Parisa Babaali, Gigi Yuen-Reed, Ryan Kather, Dario Arcos-Diaz, Michael Cherner.