Waking up to data challenges

Yuanfang Guan explains how taking part in data challenges has helped her learn new analytical techniques and creatively apply them on a variety of datasets.

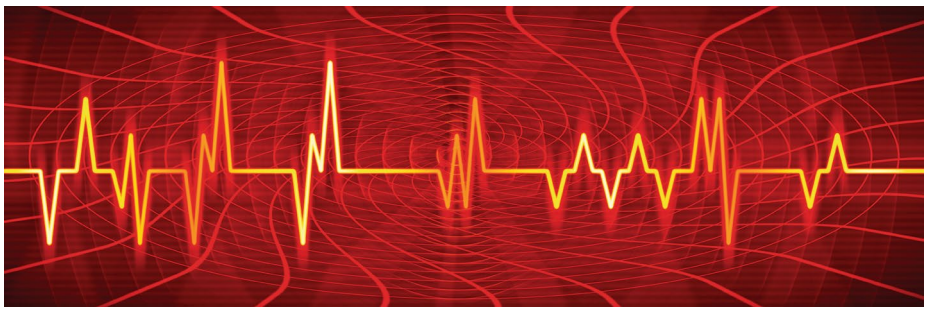
Data science challenges provide an excellent platform for scientists to dive deep into a range of problems. They encourage researchers to test their skills, quickly grasp state-of-the-art methods and improve on them — while also helping industry to identify and deploy the most accurate algorithms.

I have thoroughly enjoyed participating in (and occasionally winning) data challenges over the past five years: I took part in fifteen DREAM Challenges (http://dreamchallenges.org), which focus on biomedicine, and three Data Science Bowls (https://datasciencebowl.com), which focus on computer vision. This year, my team took part in the PhysioNet challenge (https://physionet.org/challenge), organized by the Computing in Cardiology Society since 2000, and we developed the top-scoring algorithm. These competitions have helped me to appreciate the importance of a variety of projects and identify problems that interest me. Moreover, taking part in these challenges has greatly improved my ability to learn new techniques.

This year's PhysioNet competition tasked participants with detecting sleep arousals in digital records from a combination of devices: electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG) and oxygen saturation (SaO_2), sampled for 6–8 hours. The dataset included 1,985 subjects, divided into a training set of 994 individuals and a (hidden) test set of 989 individuals. These studies aim to facilitate in-home monitoring of pathological arousals that may result from obstructive sleep disorders, or respiratory or neurological conditions.

The critical starting point when solving a data challenge, in my view, is to identify the correct branch of the technology tree — a much-discussed topic in my lab. For challenges that last 3–8 months, there is only time to explore one or two major branches of methods. Choosing the wrong branch can bear tremendous labour costs, and staying on the wrong branch can lead to failure of the project.

While exploring the problem, we observed the similarity of sleep arousal detection to scene segmentation in computer vision, except that sleep arousal detection deals with one-dimensional time series data rather than two-dimensional images. This observation points us to a branch of the technology tree: deep convolutional networks for segmentation. We turned to the fully convolutional network[1] and U-net[2] approaches that were developed in 2015 and which remain the state-of-the-art technology for scene segmentation. We expected that their adaptation to time series signals could outperform techniques compared to last-generation technologies, such as feature extraction using Fourier transforms.



Credit: JanMika/Panther Media GmbH/Alamy Stock Photo

In the PhysioNet challenge, we exclusively experimented with U-net architectures. U-net can transform an input to a binary output of the same size. This transformation is achieved by an encoding part of the deep network, which generates a high-dimensional feature representation of the input, and a decoding part, which outputs the desired segmentation map. Each decoding layer is concatenated to the corresponding encoding layer of the same size, creating a shortcut that allows information to flow through different depths of feature extraction. For the PhysioNet problem, the network encodes the time series data as a multiple-channel input (equivalent to red, green and blue colours in images) and decodes the information into a segmentation map of the sleep arousal regions.

A generic 'correct branch of the technology tree' approach is often sufficient to generate a model with decent performance. To optimize the model, however, we must incorporate it with specific biological or physical insights. Deep learning is prone to overfitting, which can be addressed by data augmentation: adding data to the training set by altering the original data. Physically meaningful augmentation is the key to many problems using deep learning. In image recognition, augmentation can be done by scaling or shearing the image, and by colour perturbation, for example. In our case, sleep arousal can be reflected by the activation of any part of the brain or any major muscle. We implemented augmentation by swapping the channels of EEG and EMG independently. Considering that an individual could breathe deeply or shallowly, a heart could beat faster or slower, and a brain could be more or less active, we further added temporal perturbations, scaling and shifting along the time axis. From the limited original dataset of 994 examples, we could use augmentation to generate an unlimited number of examples for training.

Like other data challenges, the 2018 PhysioNet challenge introduced my group to a new field and extended the technology tree available to me. I believe these challenges provide invaluable stimulation for life-long learning — an essential feature of the scientific process. ❐

Yuanfang Guan
Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.
e-mail: gyuanfan@umich.edu

Published online: 7 January 2019
https://doi.org/10.1038/s42256-018-0011-2

References

1. Long, J., Shelhamer, E. & Darrell, T. in *IEEE Conf. Computer Vision and Pattern Recognition* 3431–3440 (IEEE, 2015).
2. Ronneberger, O., Fischer, P. & Brox, T. in *Medical Image Computing and Computer-Assisted Intervention* (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer, Cham, 2015).