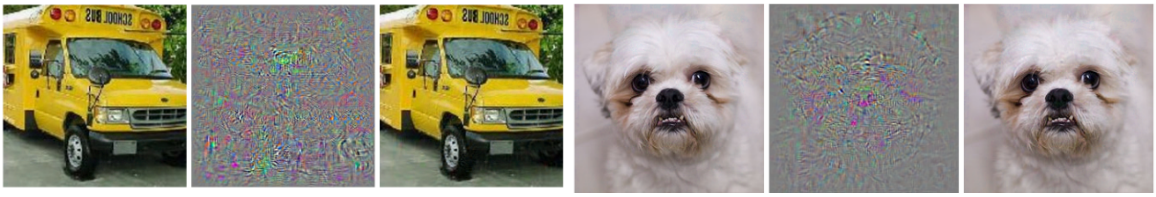
Post-doc/PhD position
Pattern Mining for Neural Networks Debugging: Application to Speech Recognition

Elisa Fromont & Alexandre Termier, IRISA/INRIA rba – Lacodam team (Rennes)
Irina Illina LORIA/INRIA nge – Multispeech team (Nancy)
firstname.lastname@inria.fr

Location: INRIA rba, team Lacodam, Rennes
Keywords: discriminative pattern mining, neural networks analysis, explainability of black box models, speech recognition.

Context:



Examples of adversarial images from [4] generated for a given DNN (here AlexNet). (Left) is a correctly predicted sample, (center) difference between correct image and image predicted incorrectly (right) as an"**ostrich**" (in both cases)

Understanding the inner working of deep neural networks (DNN) has attracted a lot of attention in the past years [1, 2] and most problems were detected and analyzed using visualization techniques [3, 4, 5]. Those techniques help to understand what an individual neuron or a layer of neurons are computing. We would like to go beyond this by focusing on groups of neurons which are commonly highly activated when a network is making wrong predictions on a set of examples. In the same line as [1], where the authors theoretically link how a training example affects the predictions for a test example using the so called "influence functions", we would like to design a tool to "debug" neural networks by identifying, using symbolic data mining methods, (connected) parts of the neural network architecture associated with erroneous or uncertain outputs.

In the context of **speech recognition**, this is especially important. A speech recognition system contains two main parts: an acoustic model and a language model. Nowadays models are trained with deep neural networks-based algorithms (DNN) and use very large learning corpora to train the huge number of DNN (hyper)parameters. There are many works to automatically tune these hyperparameters. However, this induces a huge computational cost, and does not empower the human designers. It would be much more efficient to provide human designers with understandable clues about the reasons for the bad performance of the system, in order to benefit from their creativity to quickly reach more promising regions of the hyperparameter search space.

Description of the position:
This position is funded in the context of the **HyAIAI "**Hybrid Approaches for Interpretable AI" INRIA project lab (https://www.inria.fr/en/research/research-teams/inria-project-labs)
With this position, we would like to go beyond the current common visualization techniques that help to understand what an individual neuron or a layer of neurons is computing, by focusing on groups of neurons that are commonly highly activated when a network is making wrong predictions on a set of examples. Tools such as activation maximization [8] can be used to identify such neurons. We propose to use discriminative pattern mining, and, to begin with, the DiffNorm algorithm [6] in conjunction with

the LCM one [7] to identify the discriminative activation patterns among the identified neurons. The data will be provided by the MULTISPEECH team and will consist of two deep architectures as representatives of acoustic and language models [9, 10]. Furthermore, the training data will be provided, where the model parameters ultimately derive from. We will also extend our results by performing experiments with supervised and unsupervised learning to compare the features learned by these networks and to perform qualitative comparisons of the solutions learned by various deep architectures. Identifying "faulty" groups of neurons could lead to the decomposition of the DL network into "blocks" encompassing several layers. "Faulty" blocks may be the first to be modified in the search for a better design.

The recruited person will benefit from the expertise of the LACODAM team in pattern mining and deep learning (https://team.inria.fr/lacodam/) and of the expertise of the MULTISPEECH team (https://team.inria.fr/multispeech/) in speech analysis and deep learning.

We would ideally like to recruit à 1 year (with possibly one additional year) **post-doc** with the following preferred skills:

- Some knowledge (interest) about speech processing
- Knowledgeable in pattern mining (discriminative pattern mining is a plus)
- Knowledgeable in machine learning in general and deep learning particular
- Good programming skills in Python (for Keras and/or Tensor Flow)
- Very good English (understanding and writing)

However, good PhD applications will also be considered and, in this case, the position will last 3 years. The position will be funded by INRIA (https://www.inria.fr/en/). See the INRIA web site for the post-doc and PhD wages. The candidates should send a CV, 2 names of referees and a cover letter to the **four researchers** (firstname.lastname@inria.fr) mentioned above. Please indicate if you are applying for the post-doc or the PhD position. The selected candidates will be interviewed before December 2019 for an expected start in January 2020.

Bibliography:

[1] Pang Wei Koh, Percy Liang: *Understanding Black-box Predictions via Influence Functions*. ICML 2017: pp 1885-1894 (best paper)

[2] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals*: Understanding deep learning requires rethinking generalization*. ICLR 2017

[3] Anh Mai Nguyen, Jason Yosinski, Jeff Clune: *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*. CVPR 2015: pp 427-436

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus: *Intriguing properties of neural networks*. ICLR 2014.

[5] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, Wenchang Shi: *Deep Text Classification Can be Fooled*. IJCAI 2018: pp 4208-4215

[6] Kailash Budhathoki and Jilles Vreeken. The difference and the norm—characterising similarities and differences between databases. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 206–223. Springer, 2015.

[7] Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. Lcm: An efficient algorithm for enumerating frequent closed item sets. In Fimi, volume 90. Citeseer, 2003.

[8] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. University of Montreal, 1341(3):1, 2009.

[9] G. Saon, H.-K. J. Kuo, S. Rennie, M. Picheny: The IBM 2015 English conversational telephone speech recognition system", Proc. Interspeech, pp. 3140-3144, 2015.

[10] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke : The Microsoft 2017 Conversational Speech Recognition System, IEEE ICASSP, 2018.