Task complexity shapes internal representations and robustness in neural networks

Robert Jankowski[1,2,3], Filippo Radicchi[4], M. Ángeles Serrano[1,2,5], Marián Boguñá[1,2], Santo Fortunato[4]

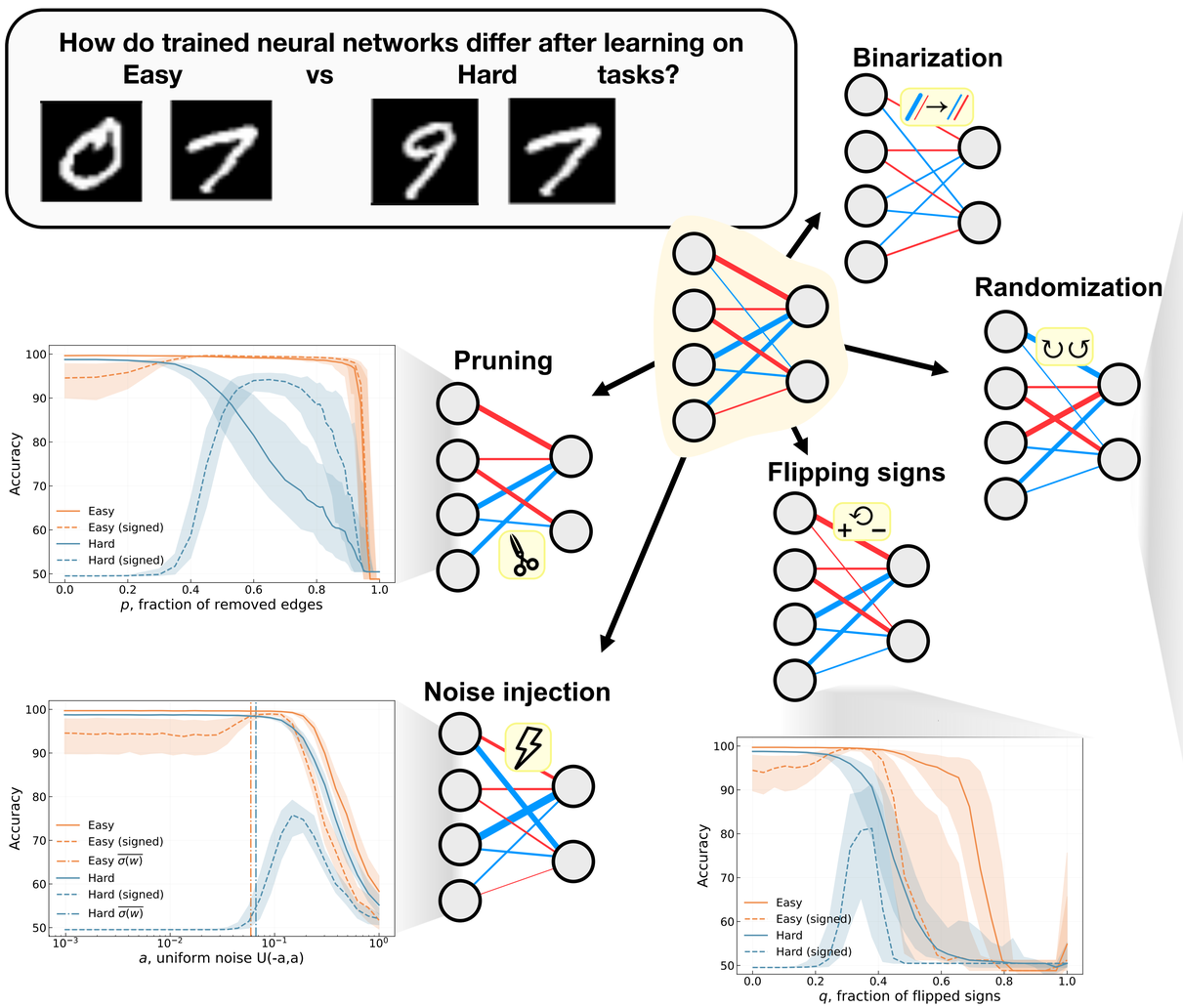
[1] Departament de Física de la Matèria Condensada, Universitat de Barcelona
[2] Universitat de Barcelona Institute of Complex Systems (UBICS)
[3] Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft
[4] Center for Complex Networks and Systems Research (CNetS), Indiana University
[5] ICREA, Passeig Lluís Companys 23, E-08010 Barcelona, Spain

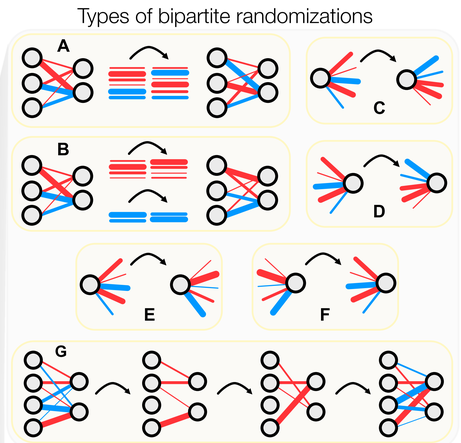
Mapping Complexity Lab
University of Barcelona | UBICS | ICREA

R.Jankowski@tudelft.nl

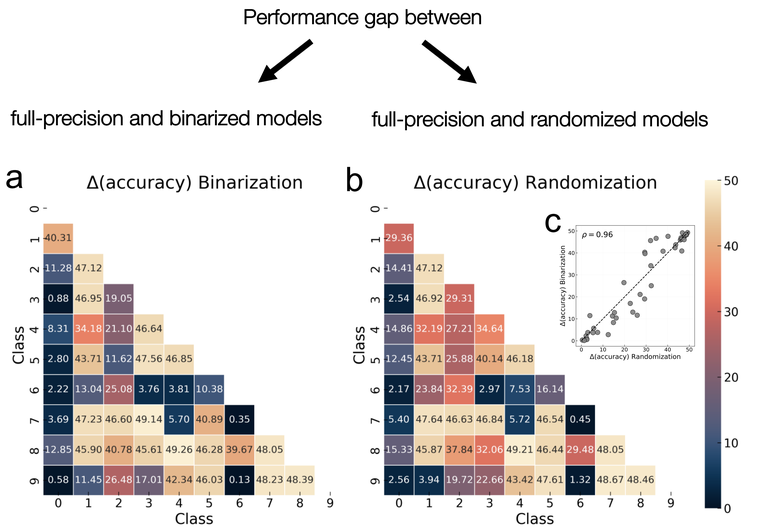
Easy and hard tasks differ in how their learned weights respond to structural perturbations

Data-agnostic probes of learned weight structure



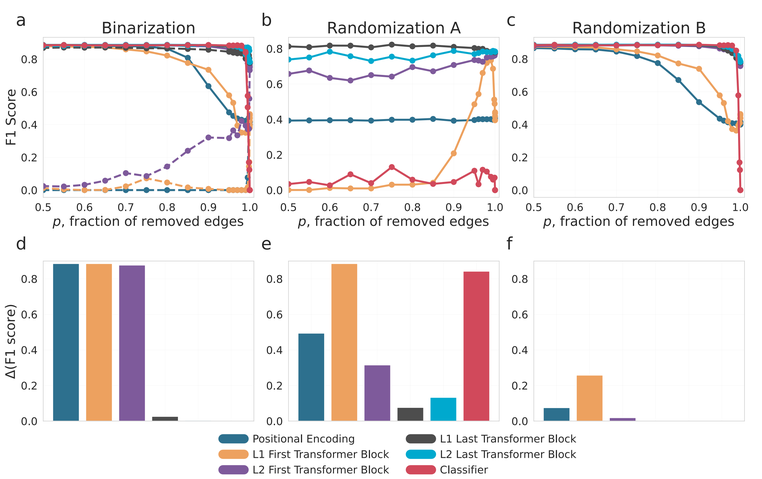


Defining task complexity



Probing transformer robustness on NER

- Robustness increases with layer depth
- Randomization A partially preserves F1 score



Acknowledgements: We acknowledge the support of the AccelNet-MultiNet program, a project of the National Science Foundation (Award #1927425 and #1927418). R.J. acknowledges support from the fellowship FI-SDUR funded by Generalitat de Catalunya. M.A.S. and M.B. acknowledge support from Grant No. TED2021-129791B-I00 funded by MCIN/AEI/10.13039/501100011033 and by "European Union NextGenerationEU/PRTR"", and Grant No. PID2022-137505NB-C22 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

[1] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[2] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612.
[3] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.