Latent Reading

by Yannis Siglidis

"Abuse of power is one of the defining features of a free society"[1]

I recently co-authored the "AI Against the Alt-Right" Twitter bot[2] in which a state-of-the-art language model (GPT-2) was trained on alt-right posts and replies from Twitter, with the purpose of generating back both posts and replies. For me, observing the behavior of such a model can allow a form of meta analysis of the alt-right parole, while isolating it from its facticity. Inspired by this I propose "Latent Reading", a research method for social sciences. In Latent Reading, instead of directly analyzing and interpreting the data-artifacts of a social entity (either individual or group),

what is analyzed and interpreted instead is their *latent representation* in a generative model that has been trained to reproduce them.

Recent advances in Deep Learning make generative modeling a much more feasible task and have motivated a research shift from studying problems of recognition to problems of generation. This has improved the expressive power of generative architectures and demonstrates their potential to accurately reproduce the statistical properties of complex forms of data, such as language. Although the required amount of (training) data increases in parallel with the evolution of deep-learning, in practice, fine-tuning a pre-trained model to a specific category of data can require significantly smaller amounts[3]. This indicates that latent reading could potentially become a low-resource interdisciplinary task.

Latent Reading draws from those studies of both social or natural complex systems (from sociology to earth-sciences), where research is not presented on observations made from a system under examination, but rather from its computer simulation[4]. In this case the research objective is not to analyze the data-output of such a system, but instead to understand how a learning system has learned to reproduce it, either by analyzing samples of its generated outputs, or by interpreting its trained architecture. Moreover, due to its nature this modeling technique is indifferent to the facticity of the given data and allows the

research findings to be posed only in terms of their latent representation (and not the subject itself). Another potential benefit of this method is that it limits interaction with the subject to that of data collection (and is thus absent for data trails). Last but not least, the fairness of such architectures is an open research problem that is being increasingly studied and addressed by respective scientific communities[5].

1. https://twitter.com/radicaldumb/status/1379032768680693764.

2. https://twitter.com/radicaldumb.

3. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al., "Language Models Are Few-Shot Learners," arXiv.org (June 1, 2020), arxiv.org/abs/2005.14165v4.

4. Eric Winsberg, « Computer Simulations in Science », *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), plato.stanford.edu/archives/win2019/entries/simulations-science/.

5. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning," arXiv.org (September 17, 2019), arxiv.org/abs/1908.09635.