MACHINE LEARNING FOR GENOME ANONYMIZATION

1 PROBLEM

Technological advance has lead to a growth in available health data, far outpacing the development of methods capable of harvesting that knowledge (1). Genomes are one of the types of information being generated in vastly larger numbers than before. Yet, while publicly available data sets are strictly separated from information about the individual, complete deidentification of the individual remains a highly non-trivial task with many adversarial tactics still finding success (2). For instance, from date of birth, zip code and gender 87 % of the population in the united states can be uniquely identified (3). Therefore, effective prevention of reidentification of full genomes is a prerequisite to make use of the immense value in genomic data.

2 APPROACH

One approach would be the replacement of a dataset by a statistically identical one before public release, ensuring the security of contributing individuals.

Recent results have shown the applicability of deep learning in health data (4). For instance, *generative adversarial networks* (GAN) which have been successful in creating new examples indistinguishable from the training set (5). In this method, we consider n examples $\boldsymbol{x} = \{x_1, \dots, x_n\}$ generated by unknown probability distribution $\mathcal{P}_x$. Next, we introduce a generator G creating examples x_g that attempt to imitate the data in $\boldsymbol{x}$. The generator is confronted with a discriminative network D that given sets of $x_g \sim G$ or $x_i \sim \mathcal{P}_x$ tries to determine the source of the image (6). During the training process both D and G are parametrically optimized to increase performance on their respective tasks. I prepared a repository `https://github.com/LukasFrankenQ/genome_GAN` and established basic functionalities on the MNIST dataset, the results of which are shown in Figure 1. The left side exhibits microscopic properties in resulting generated examples, some of which are indistinguishable from examples $x_i \sim \mathcal{P}_x$, while some others are clearly artificially generated. The right side evaluates a macroscopic property in the average over 1000 generated examples versus examples drawn from the dataset. We find similarities in the overall shape but also observe discrepancies, particularly in overall smoothness.

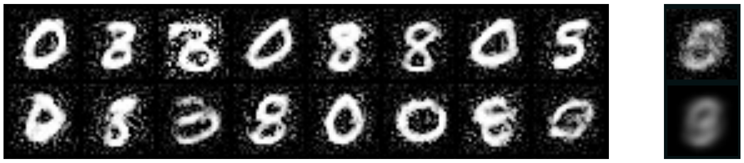

Figure 1: Example of retained properties during data generation for MNIST (7). *(Left)* Examples generated by G after 200 epochs of training on a basic network architecture. *(Right)* Average of 1000 examples *(Top)* randomly sampled from G, *(Bottom)* randomly drawn from the MNIST dataset.

This toy example aims to shed some light on how statistical properties are retained or lost by GANs. Naturally, the performance of the exhibited networks is not comparable to the fields state of art, as for example shown on this website (here) and much better results can be expected for the MNIST case with the appropriate computational power. However, the complexity in the human genome vastly diminishes the complexity in handwritten digits giving rise to the following questions:

- Which statistical properties are retained or lost as the complexity of the original dataset is increased?
- How can full genomes be transformed into a feasible training-set?

References

[1] Nabil R Adam, Robert Wieder, and Debopriya Ghosh. Data science, learning, and applications to biomedical and health sciences. *Annals of the New York Academy of Sciences*, 1387(1):5–11, 2017.

[2] Suyash S Shringarpure and Carlos D Bustamante. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5):631–646, 2015.

[3] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[4] Sara Khalid, K Berensci, M Ali, Peter Rijnbeek, and Daniel Prieto-Alhambra. Deep learning for drug utilisation research: Identifying features within a population of anti-osteoporosis drug users. 2018.

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. 2010.