The Doppelgänger Effect in Machine Learning

Chutong Deng

1. Doppelgänger effect: What is it, and how does it affect the machine learning model?

With their ability to learn patterns from datasets and make predictions accordingly, machine learning (ML) models are being increasingly used in biomedical data science. In drug design, ML models can significantly reduce the time spent on discovery and testing by quicky shortlisting effective drug candidates [1]. However, while the use of ML in biomedical data science has shown promise in recent years, there are concerns about the reliability of the validation results of the models due to the existence of the doppelgänger effect.

The doppelgänger effect in ML refers to the phenomena where a model trained on a dataset attains excellent performance on a separate, but related dataset due to the high degree of similarity between the training and testing data [1].

The doppelgänger effect can result in overfitting and subpar generalization performance. Although evaluation accuracy may appear to be high due to the presence of data doppelgänger in training and testing sets, this might lead to inaccurate conclusions about the performance of the models. Models that have only been trained on a single dataset might not be able to adapt adequately to brand-new, untested data. This can occur when a model is trained on a dataset that does not accurately represent the population to which it will be applied or when a model is trained on a dataset with a different feature distribution than the new dataset.

In order to improve the robustness and generalization capabilities of ML models, it is thus critical to look for methods that can check for and avoid the doppelgänger effect.

2. Is doppelgänger effect unique to biomedical data?

The doppelgänger effect may be observed in the biomedical data science in a number of tasks including risk prediction, disease diagnosis, and drug discovery.

The great degree of resemblance across various patient groups is one factor contributing to the doppelgänger effect in healthcare and biological data. For instance, people with the same condition sometimes have comparable symptoms and medical histories, and similar forms of data, including test results and prescription lists, may be found in electronic health records. Machine learning models will be more likely be developed and evaluated on datasets that contain data doppelgänger in biological data science. In several fields of bioinformatics, including protein function prediction and drug development, the Doppelgänger effect is evident. Proteins with similar sequences are presumed to have comparable functions in protein function prediction, although the method may not accurately predict functions for proteins with less similar sequences but similar activities. Models that predict the biological activity of molecules in drug discovery based on their structural characteristics make the assumption that molecules with similar structures will have similar activities, but this premise makes it impossible to distinguish between poorly trained models and well-trained ones [1]. This may give the appearance that forecasts are accurate when they are not.

In order to identify chest x-ray images for pneumonia, Wang et al. [2] utilized a deep learning model. They discovered that the model performed well on a separate dataset of chest x-ray images from a different hospital. However, the author stated that the high degree of similarity between the two datasets, which may have resulted in poor generalization performance when applied to new datasets that are not similar, may have contributed to the model's ability to transfer well to new data.

Although the doppelgänger effect is common in biomedical data science, it is not limited to biomedical data. The doppelgänger effect can also be seen in the field of Natural Language Processing (NLP) and Computer Vision (CV).

The doppelgänger effect in text generation occurs when a model produces text that is remarkably similar to the input it was trained on. As a result, the model may produce content that is monotonous, illogical, or biased in favor of the training set. Based on a dataset of Amazon product reviews, Radford et al. [3] refined a pre-trained language model and discovered that the model produced text that was very comparable to and closely mirrored the input reviews, even duplicating phrases and sentences verbatim.

A model was employed by Shrivastava et al. [4] and tested on real-world images after being trained on a collection of synthetic images. Due to a discrepancy between the distributions of synthetic and real images, they discover that learning from synthetic images may not produce the intended performance. In order to solve this problem, they employ adversarial training, in which the model is trained on both synthetic and real-world images in order to enhance its performance on the latter.

3. How to check and avoid doppelgänger effect in practice?

The doppelgänger effect can be a significant concern for machine learning models, as it can lead to overfitting and inadequate generalization performance. Therefore, it is important to detect and prevent the doppelgänger effect in real-world practice.

Several methods have been proposed to identify data doppelgängers [1], including ordination methods, embedding methods, and measures such as dupChecker [5] and the pairwise Pearson's correlation coefficient (PPCC) [6].

There may be several possible ways to reduce the effect of data doppelgängers.

The doppelgänger effect may be mitigated by feature engineering and data splitting techniques based on the specific context of the data being analyzed, such as identifying and removing data doppelgängers, using different cell types to form the training-evaluation pair, and splitting training and test data based on individual chromosomes. However, these solutions have limitations since they depend on prior knowledge and high-quality contextual data, or they could not work well with small data sets that have a large number of data doppelgängers [1].

The usage and interchange of biological data among different institutions and research groups has become easier due to electronic health records and other sizable databases. The possibility of finding less comparable data increases, which raises the likelihood of constructing a more generalized model. Other possible approaches include using ensemble methods, which involve

training multiple models on different datasets and combining their predictions to make a final prediction, and transfer learning, which allows a pre-trained model on one task to be used for a different yet related task.

4. Conclusion

The doppelgänger effect commonly occur in machine learning, especially when studying medical and healthcare data. It happens when the training and test sets of data are remarkably similar. The doppelgänger effect is a difficulty in NLP, CV, and other machine learning domains, in addition to being ubiquitous in biological data.

The doppelgänger effect can cause falsely-positive impressions about a model's performance, which is a serious risk for machine learning models. To detect and reduce data doppelgängers, a number of strategies have been devised, but each has substantial limitations and drawbacks. There is still much to discover about the doppelgänger effect, making it a promising area of study.

References:

[1] Wang LR, Wong L, Goh WW. How doppelgänger effects in biomedical data confound machine learning. Drug Discovery Today. 2021 Oct 28.

[2] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 2097-2106).

[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

[4] Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R. Learning from simulated and unsupervised images through adversarial training. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 2107-2116).

[5] Sheng Q, Shyr Y, Chen X. DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis. BMC bioinformatics. 2014 Dec;15(1):1-3.

[6] Waldron L, Riester M, Ramos M, Parmigiani G, Birrer M. The Doppelgänger effect: Hidden duplicates in databases of transcriptome profiles. JNCI: Journal of the National Cancer Institute. 2016 Nov 1;108(11).