Review of doppelgänger effect and features analyses

Abstract

The main purpose of this report is to summarize the characteristics of doppelgänger effect, analyze the possibility of doppelgänger effect in other fields that it may occur in Speech recognition and AI painting, and try to propose ways to avoid it, like using other dimensions of the data, correlation analysis, and random noise.

Key words: Doppelgänger effect, machine learning

Introduction

In machine learning models for drug development, researchers have observed that independently derived training and test sets can still produce unreliable validation results, that is, the performance on the validation data set is good regardless of how well the model is trained, Gives people the illusion of overestimating the training results. This is due to the high similarity between the training set and the test set. Li RongWang et al. proposed that there is a large amount of data doppelgänger in biomedical data. And the method of identifying doppelgänger effect is analyzed. Researchers are working hard to analyze the mechanism by which the doppelgänger effect occurs, develop methods to detect the doppelgänger effect, and mitigate the impact of the doppelgänger effect.[1]

The doppelgänger effect has also emerged in other related fields. An example is the doppelgänger effect in the Transcriptome Atlas database. Researchers often share or reuse specimens in future studies. Duplicate expression profiles in public databases will affect reanalysis and lead to doppelgänger effect. Due to privacy protection, the cancer data does not point to the name of the source, but it may be adopted by multiple data sets. For example, Levi Waldron et al. found that about 17% of the records in the ovarian cancer database were not unique. The original system of identifying duplicate data will not identify these data.[2]

Li RongWang et al. [3]pointed out that although it is difficult to remove data doppelgängers directly from the data, it is possible to use metadata as a guide to conduct careful cross-checking, perform data stratification, perform divergence validation involving as much as possible, etc.

On this basis, this article will explore the possibility of doppelgänger effect appearing in other fields, and try to find out the means to avoid doppelgänger effect.

Doppelgänger effect in other fields

This paper argues that the doppelgänger effect may also appear in other fields. The source of doppelgänger effect is similar training set and test set, so it is easy to appear doppelgänger effect in biomedical data, which has small sample size, low definition, small difference between samples, and difficult abductive reasoning. This paper suggests several areas where the doppelgänger effect is likely to emerge.

Speech recognition data. In the recording data, there are a lot of noise and similar fragments, and it is difficult to obtain a large amount of data sets for the recognition of some uncommon languages (such as dialect collection and recognition projects). That is, the training set and the test set may have a high degree of similarity, leading to overestimation of the training results.[4]

AI painting. Art is highly abstract, and two completely different emotional or metaphorical descriptions are easily attributed to the same label by the Internet.[5] As in the field of drug discovery, proteins with similar sequences are inferred to be from the same ancestral protein. Moreover, the

works of human painters as a training set are limited. If you want to develop an AI that imitates a certain style of painting, you will inevitably encounter many repeated or similar data sets.

Ways to avoid doppelgänger effect

There are several ways to avoid doppelgänger effect.

Cross-check using other dimensions of the data as a guide. Such as Li RongWang et al. use metadata in RCC to construct negative and positive cases.

Use Pearson correlation analysis to perform as many correlation checks on the data as possible, develop a doppelgänger effect checker, to perform the possibility of "effect inflation" after the test is expected.

It is also possible to reduce the doppelgänger effect by adding random noise to the data and randomly batching the data for training.

Conclusion

This article summarizes the current research on doppelgänger effect, and proposes areas where doppelgänger effect may exist, and summarizes and analyzes methods to avoid doppelgänger effect.

References

[1,3,6] Wang L R, Wong L, Goh W W B. How doppelgänger effects in biomedical data confound machine learning[J]. Drug Discovery Today, 2021.

[2] Waldron L, Riester M, Ramos M, et al. The doppelgänger effect: hidden duplicates in databases of transcriptome profiles[J]. JNCI: Journal of the National Cancer Institute, 2016, 108(11).

[4] Çayır A N, Navruz T S. Effect of dataset size on deep learning in voice recognition[C]//2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2021: 1-5.

[5] Santos I, Castro L, Rodriguez-Fernandez N, et al. Artificial neural networks and deep learning in the visual arts: A review[J]. Neural Computing and Applications, 2021, 33: 121-157.