Q3 Report: The doppelgänger effect

Executive Summary

The primary purpose of this report is to review the doppelgänger effect [1] occurring in biomedical machine learning studies. Variously appearing in the natural image understanding tasks and biomedical analysis with machine learning, the doppelgänger effect misleads the optimization process of the AI models. In specific tasks like classification, the doppelgänger effect may cause a discordant issue in their interpretability analysis and affect the model's generalizability. By analyzing some cases, this report summarizes the reason and result of its occurrence. Some ideas are reviewed and developed to alleviate their influence on the machine learning process of biomedical data. A specific case with AI-onsite pathology image analysis is analyzed, exploring the clinical sampling process possibilities.

Introduction

Since the new millennium, artificial intelligence research has boosted and achieved inspiring results in comprehensive applications. Researchers have conducted extensive research on machine learning design as an automated attempt to model various data features. Based on big-data studies, many machine learning methods are introduced to extract features and identify abstractive patterns [2]. A typical classification task requires three processes: a training process in which a massive dataset is used to supervise the machine learning model to constrain, a validating process in which a representative dataset is used to select the best-performed model, and lastly, a testing process which an independent dataset is used to measure the performance of the trained model. By the three-step design of training, validating, and testing, the model is reported with a series of criteria that quantitatively measure the model's understanding of the dataset [3]. This method enables the researchers to optimize the structure or the training strategy of machine learning models to obtain a higher overall result for its better applicative performance.

However, the traditional three-step optimization process may be misled by the inner co-relations of the dataset. The bias and the distributions of datasets may cause severe deterioration in the application of machine learning models. Many studies are carried out to explore a better machine learning method that can be more robust despite the perturbation of datasets. The methods to alleviate the dataset distribution issue include optimization studies

based on loss design [4], introducing the pre-knowledge design in model structure [5], and using the representation learning method before the training process [6], etc. Another trend in understanding the impact of datasets is to focus on relevant information that some samples may share [1-3]. The distributions and co-relation features of specific samples in the dataset can prominently affect the optimization process, drawing the biomedical data science community [1] as the area needs reliable methods. Remarkably, the concept of data doppelgänger underlines that specific samples may mislead results in the measuring process of the machine learning models. The data doppelgängers appear when similar data (or duplicated ones) are divided into different datasets, and therefore they may misguide the model to report output relating to their co-relation. When these datasets are used in the training and validating process, the co-relation of specific similar data makes the performance measurement unreliable [1]. Although the influence of data doppelgängers raises several attentions, it remains un-explored as few studies can be viewed. As a pivotal issue in machine intelligence's safety and interpretability field, more understanding of data doppelgängers needs to be reviewed by examples and solutions.

1. The data doppelgängers in nature image analysis

As a typical similar data pair, twins' faces are obvious samples in the human face dataset, leading to data doppelgängers occurrence [7]. In the cases that twins are taking into datasets of training and validating or the cases where a subject was taken photo more than once, the identical duplicate data lead to data doppelgängers. Furthermore, many similar image samples may exist as a particular object may be taken photos many times with different surroundings (for example, a specific type/brand of bottles may exist on many occasions) and reserved in a separate dataset. Since the size of nature image analysis is generally much bigger, the data doppelgängers in nature image understanding may not as identical as the same occurrence in medical datasets. The texture and co-occurrence characters may be another possibility where the data doppelgängers should be monitored. A similar co-occurrence may be understood with the example of the ostrich, which is one of the birds that can't fly, such character leads to the background co-occurrence without sky or cloud. In a hypothetical fine-grained classification task to distinguish birds, the non-ostrich birds are with background instance of the sky and the ground while only the ostrich has the none-sky background. Such co-occurrence may be linked to data doppelgängers.

2. The data doppelgängers in biomedical machine learning analysis

Following the findings in [1], several examples address this issue in machine learning research using biomedical data. Firstly, due to the similarity of specific samples, data doppelgängers can be viewed among various modalities. In [8], a chromatin interaction prediction system was reported to be overstated in the results as similar samples are shared in both training and validation datasets. In [9], Levi Waldron et al. analyzed the duplicate expression profiles in public databases, leading to the doppelgänger effect. Secondly, in several explored fields of bioinformatics, the doppelgänger effect exists along with the specific characters of the samples. In the proteomics example offered in [1], two different proteins with similar sequences may lead to data doppelgängers in predicting functions since their similar sequence indicates the inheriting of their functions from the same ancestral protein. The authors explored the doppelgängers effect with pairwise Pearson's correlation coefficient (PPCC) design on renal cell carcinoma (RCC) proteomics. By building training and validating datasets from different groups of data pairs, the PPCC data doppelgängers are defined and analyzed. Furthermore, the related pre-knowledge may also cause the doppelgängers effect in a machine learning study predicting the activities of molecules [10]. The authors classified similar molecules with similar activity into training and validation sets, which may generate undertrained models reporting inflated results based on similar features. Lastly, in certain machine learning methods, including representation learning [6,11], etc., the doppelgänger effect may occur as the data are grouped with sharing features.

3. The strategies in identifying data doppelgängers and alleviating the effect

A straightforward strategy to identify data doppelgängers is to view the samples feature in a reduced-dimensional space [1]. However, for many cases, the data doppelgängers are not distinguishable, requiring more sophisticated methods, including dupChecker [12] and PPCC [13]. In the methods given in [1], to obtain the clear-cut scenarios using sample pairs from different categories, the researchers build negative cases, validate cases and positive cases where data doppelgängers do not exist, may exist, and promisingly exist. Then, based on the PPCC distribution of the valid scenario against its negative and positive counterparts, the PPCC data doppelgängers are identified. To mitigate the effect of data doppelgängers, a possible idea is to remove these samples so the model can learn the identical features preventing this potential perturbation. Another attempt in [1] is to remove the correlated

variables linked to data doppelgängers effects; this attempt achieved little results due to the complexity of their inner relations. Reducing the sample's inner co-occurrence may be a possible way to alleviate the effect of doppelgängers. A detailed strategy can be introduced by applying the MIL method and splitting the samples into bags with an overall weak label.

4. A case analysis with the pathology images

Since the doppelgänger effect is a pivotal issue in biomedical analysis and has been wildly observed in different modalities, the image analysis of pathology samples may share the same problem. In the pathology image analysis, the privacy agreement and the data scarcity made the labeled dataset relatively small in general, which is easy to be misled by the inner co-relations of the samples. Furthermore, the cytopathology samples from a patient often share the same background character as they are obtained and stained simultaneously. In the machine learning tasks with cytopathology samples, the similarity between samples is wildly seen [14]; only fine-grained features and global distributions distinguish the different categories. Therefore, it can severely influence the models if the doppelgänger effect was neglected in the validating process, which may hinder the performance of computer-aided diagnosis systems on the clinical samples.

With the cutting-edge Transformer and the multi-instance learning (MIL) method, a possible solution to alleviate the doppelgänger effect can be described as follows: Firstly, split each batch of image samples into several bags with a series generated patch-level label. Secondly, regroup the patches into shuffled bags of the batch and generate the bag-level soft label for every bag in the same batch by aggregate operation. Thirdly, train the MIL model with shuffled sample-batch each epoch. With the shuffled bags, the Transformer model is forced to learn the spatial difference of each patch instead of the similar global samples. Such exploration may optimize the generalization difficulties that the Transformer based models share in the small dataset conditions.

Conclusion and Recommendations

In conclusion, the data doppelgängers occur in various machine learning tasks, and as AI-based applications flourish, the impact of the doppelgänger effect continues to attract researchers' attention. Several studies defined the data doppelgängers as similar samples or the duplicated ones in the training and validating dataset, which may mislead the

measurement of the trained models. By identifying the data doppelgängers in various biomedical and nature image understanding tasks across different modalities, it is reviewed with several examples. Following the research in [1], the data doppelgängers can be identified quantitatively. The introduction of PPCC data doppelgängers sheds light on the application of a reliable method for alleviating the inner co-relation issue.

By taking a particular example in the pathology image analysis, the MIL strategy may be applied to solve the similarity between the samples by splitting the samples into image patches and supervising the MIL model with aggregated weak labels. The proposed method may alleviate the data doppelgängers problems with the PPCC method and mitigate the generalization shortcoming of the Transformer models in the small dataset.

Reference

[1] Wang, Li Rong, Limsoon Wong, and Wilson Wen Bin Goh. "How doppelgänger effects in biomedical data confound machine learning." Drug discovery today (2021).
[2] S.Y. Ho, K. Phua, L. Wong, W.W.B. Goh, Extensions of the external validation for checking learned model interpretability and generalizability, Patterns 1 (2020) 100129.
[3] Waldron, Levi, et al. "The doppelgänger effect: hidden duplicates in databases of transcriptome profiles." JNCI: Journal of the National Cancer Institute 108.11 (2016).
[4] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.
[5] Li, Kunpeng, et al. "Tell me where to look: Guided attention inference network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[6] Kang, Hengyuan, et al. "Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning." IEEE transactions on medical imaging 39.8 (2020): 2606-2614.
[7] Röttcher, Alexander, Ulrich Scherhag, and Christoph Busch. "Finding the Suitable Doppelgänger for a Face Morphing Attack." 2020 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2020.
[8] Cao, Fan, and Melissa J. Fullwood. "Inflated performance measures in enhancer–promoter interaction-prediction methods." Nature genetics 51.8 (2019): 1196-1198.
[9] Waldron, Levi, et al. "The doppelgänger effect: hidden duplicates in databases of transcriptome profiles." JNCI: Journal of the National Cancer Institute 108.11 (2016).
[10] Muratov, Eugene N., et al. "QSAR without borders." Chemical Society Reviews 49.11 (2020): 3525-3564.
[11] Zhang, Changqing, et al. "Generalized latent multi-view subspace clustering." IEEE transactions on pattern analysis and machine intelligence 42.1 (2018): 86-99.
[12] Sheng, Quanhu, Yu Shyr, and Xi Chen. "DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis." BMC bioinformatics 15.1 (2014): 1-3.
[13] Waldron, Levi, et al. "The doppelgänger effect: hidden duplicates in databases of transcriptome profiles." JNCI: Journal of the National Cancer Institute 108.11 (2016).
[14] Zhang, Tianyi, et al. "MSHT: Multi-stage Hybrid Transformer for the ROSE Image Analysis of Pancreatic Cancer." arXiv preprint arXiv:2112.13513 (2021).