Doppelganger Effect in Machine Learning Models

Weiyu Han

January 16, 2023

Abstract

This report explores the prevalent existence of the Doppelganger effect in areas other than biomedical data science and shows possible ways to detect the Doppelganger effect and lists some solutions to resolve them. Based on the analysis of aforementioned solutions, we further propose robust feature selection as pillar strategy. This report discusses how robust feature selection in the biomedical field can reduce the impact of the Doppelganger effect and the relevance of the Doppelganger effect to the alignment problem of artificial intelligence.

Keywords: Doppelganger Effect; Machine Learning; Biomedical Data

1 Introduction

- Doppelganger effect usually occurs when a machine learning model is trained on data with duplicate or highly similar samples(1), and therefore leading to overfitting and poor generalization of the model to new data.
- Due to the high volume of repeated or similar sequences from natural evolution, the repeatability of the training set and test set of the model will inevitably be relatively high(2), which can cause inflated generalization test results. Doppelganger effect is ubiquitous in biomedical machine learning problems. In fact, if there are large functional changes caused by small important structural differences, machine learning models usually perform poor to identify them well(3). This will lead to insufficient training of the model, because it is trained by a lot of less important information. As a result, the significant information might be weakened. This kind of model may have overfit problems(4) or face the problem of being unable to accurately predict Out-of-distribution (OOD) data, which arises when a machine learning model sees an input that differs from its training data, and thus should not be predicted by the model(5), which will limit the generalization ability of the model.
- Furthermore, we propose that the Doppelganger effect does not only exist in biomedical data but exists in many machine learning tasks (see Section 2). Indeed, biomedical data has common problems such as many repetitive sequences caused by evolution and mutation(6), however, such effect broadly exists among various types of data due to the similarity between training set and validation set is not rare.
- This report summarizes the existing literature and propose some potential feasible solutions on how to avoid the Doppelganger effect in practice, from feature selection to model training, in order to avoid the impact of Doppelganger effect on health and medical science machine learning models to the greatest extent. Besides, the report will propose a possible connection between Doppelganger effect and alignment problems in AI safety.

2 Prevalent Doppelganger Effect

- Dataset shift, the problem that the test set and train set do not satisfy the same distribution, is prevalent in machine learning settings(7). The phenomenon of dataset shift also includes three problems mainly. The first one is covariate shift(i.e., in the image recognition task, a face without a mask is input during training, but there are a lot of face data with a mask in the prediction, which makes the machine unable to recognize it accurately). The second one is prior probability shift, which referring to the shift caused by the difference between the label distribution of the training data and the predicted data. And the last one is concept drift, such as the most typical farmer theory. The Doppelganger effect is a typical case of the data shift phenomenon, which is ubiquitous in many machine learning models, although we emphasize its serious impact in biomedical machine learning problems. In fact, in other fields, the Doppelganger effect also exists widely.
- A typical example of Doppelganger effect is the use of machine learning models to identify spam(8). The spam classifier users usually find that its real accuracy is much lower than the claimed accuracy. In detail, the machine learning models sometimes miss spam that humans can easily spot, and sometimes include documents from important customers in the spam. This is one of the manifestations of the Doppelganger effect. The reason

that may cause this phenomenon is that the developers of machine learning models may select many data to train their model in a accurate way. However, there are many types of spam in practical use, and even some spam will bypass the shielding according to the classification strategy of the machine learning system. The high score classifier gets on validation set may be inflated by the similarity of train set and validation set. In other words, the Doppelganger effect occurs.

- Another prevalent example of Doppelganger effect is the use of AI in image recognition. The researchers attempted to build a machine learning model to identify pet dogs and selected many pictures of pet dogs(9). Most of these pictures are taken on the lawn (pet breeders may prefer to take pictures when their pets are playing on the lawn), which will lead to a certain similarity in the selection of the training set and the validation set (grass). Ergo, the trained model can recognize pet dogs on the lawn very well, but when it recognizes pet dogs on the road or in the water, it will produce poor recognition results. When checking the trained model, the model is observed to assign a greater weight to the image feature ”grass”. However, there is no causal relationship between the two propositions ”whether the background of the picture is grass” and ”whether it is a pet dog”. The model gets this ridiculous result partly because the Doppelganger effect.
- We therefore believe that the Doppelganger effect is a pervasive problem throughout the field of machine learning. In summary, the above two examples have a similarity, that is, the data and features selected when training the model cannot fully cover the complex situation that the model may encounter when it predicts. Therefore, we have come to a revelation: features and data are always the biggest challenges for machine learning, which also determines the upper limit of machine learning model predictions. The Doppelganger effect should be taken seriously in the field of machine learning.

3 How to Identify Doppelganger Effect

- Doppelganger effect curtails generalization performance of machine learning models, and unfortunately, in many cases, we do not have a good way to identify the Doppelganger effect and define well which pairs of Doppelganger effect is acceptable in model training. However, we got some inspiration from ”How doppelgänger effects in biomedical data confound machine learning.” by Wang(1). Wang proposed the method of pairwise Pearson’s correlation coefficient (PPCC) before judge the score of models on validation set. By calculating all the PPCC between different data pairs and determining the possible PPCC range of the data pairs with Doppler effect may occur, Wang listed all these data as Doppelganger data. Besides, they find that isolating all Doppelganger data in train or validation set can avoid Doppelganger effect as much as possible (although it may produce overfit or winner-take-all situation). However, this provides a feasible solution for data sets with sufficient sample size, which is deleting one of the sample pairs with PPCC Doppelganger effect from the data set.
- Another feasible approach is to use the Population Stability Index (PSI)(10). Calculate the PSI of the train set and the validation set to measure whether the distribution ratio of the validation set and the train set is consistent, so as to evaluate the similarity between the two. At the same time, the PSI of the sample set, which may be encountered in the future, and train set is calculated to measure the similarity between the train set and the test set. If the former PSI is obviously larger, it proves that there may be data Doppelganger. However, this method can only be used as a potential solution for qualitative analysis, and it cannot accurately measure whether the data Doppelganger is acceptable, and therefore has relatively large limitations.

4 How to Mitigate Doppelganger Effect

- After identifying the possible Doppelganger effect, we therefore propose some solutions to avoid the influence of the Doppelganger effect on machine learning as much as possible. At present, there are some prevalent and old-fashioned methods in the field of machine learning to avoid such problems. Commonly used methods include data preprocessing(11), data augmentation(12), ensemble (bagging, boosting, stacking, etc.)(13), setting an independent verification set, increasing data diversity, etc.
- These methods are diversely applied in different stages of optimizing machine learning models. However, different practical rules were developed in different situations. For example, in the data preprocessing stage,

some clustering methods (such as KNN)(14) can be used to gather similar sample clusters together, and each separated sample cluster can be used as a separate sub-data set to train and verify the model. Although the method of operation is similar, the specific details of how to cluster and how many clusters are needed still depend on the research question. Another example is data augmentation. It is often used when it is difficult to increase the amount of data. By getting as much representation from the original data as possible without increasing the metadata, it can improve the quantity and quality of metadata features and improve the robustness of the model. Similarly, different practical methods are developed in different field when solving computer vision (CV) problems. In medical CV problems, repeated cell pictures often appear. To avoid overfit problems in machine learning models, the pictures in train set are stretched, rotated, translated, and reversed to prevent overfit. But sometimes the data augmentation can be counterproductive, such as in the field of face recognition(15), there is no so-called ”inverted” face, so this kind of data augmentation is an invalid operation and may even reduce the accuracy of the model. Therefore, although we hope to hedge against the impact of the Doppelganger effect, it is also crucial to choose the right method for the right field.

- In summary, traditional solutions is only able to alleviate Doppelganger effect considering its complexity. But it is still necessary to have a clearer understanding of the analysis problem itself, so that an appropriate method can be selected to reduce the Doppelganger effect and improve model. Hence, this leads us to pay more attention to the characteristics of specific problems, that is, to return to the feature selection problem of machine learning.
- Besides the general and non-exhaustive plans mentioned before, we consider the recent advance on rethinking of machine learning models in molecular modelling and find another example of Doppelganger effect. It is worth noting that in the world of biochemistry, more data are sequences and molecular/protein structures, so how to analyze sequences, graphs (molecules can be modeled as 2D/3D graphs), point clouds (protein structure information) Robust feature extraction remains an issue. And the prevalent use of machine learning models is to simply infer the functional similarity by the inherited similar structure. However, recent research challenges this method by considering that it cannot achieve a good discrimination effect when encountering activity cliffs(16). Recently, Tilborg et al. (16) propose to solve the activity cliffs. Tilborg D extracted three similar structures in the chemical formula that greatly affect the molecular function: Substructure similarity, Scaffold similarity, and SMILES similarity as considerations for model training. This method of extracting key structures is a solution that medical science researchers can use to mitigate the Doppelganger effect. Besides, they also found that traditional chemical descriptions were sometimes more effective than graph-based methods. Therefore, how to select truly useful and robust features will always be a problem that biological/chemical data science will face. In summary, due to the particularity of biochemical data (as mentioned above), it will be very valuable to extract key features and perform feature engineering exploration, which will also be the long-term development direction of biochemical data science in the future.

5 Doppelganger Effect and Alignment Problems in the AI Safety Field

- The alignment problem(17) is likely to appear after the traditional artificial intelligence research field develops to a certain level of complexity. For most of the current problems, we can accurately communicate to the machine what it needs to do based on our own experience or desired results. However, when it comes to some complex tasks, the designer cannot sort out every detail or the AI has omitted the critical information during the training process, and the designer may not be able to make the AI achieve the desired training result (the training results of AI do not exactly match the direction expected by the designer)(18). An alignment problem arises at this time.
- The Doppelganger effect is, in essence, also an alignment problem, and biological data scientists hope that trained machine learning models can accurately identify the functional groups or structures that are responsible, even if these features only occupy a very small weight in the training set. But in fact, the overfit problem caused by the Doppelganger effect may make some originally unimportant features play an important role in the classifier, which is a realistic manifestation of the AI alignment problem.
- Combining the alignment problem and the Doppelganger effect, one of the major challenges currently faced by machine learning is the feature selection. In biomedical data science, there has been a lot of recent research on how to select more representative and robust features. These studies include algorithmic improvements,

such as the Nagpal(19) who optimized the existing evolutionary computation (EC) algorithm and developed the gravitational search algorithm (GSA). GSA can reduce 66 percents of features and improve prediction accuracy by collaborating with KNN algorithm. The Doppelganger effect can be effectively alleviated by reducing the amount of irrelevant or redundant features with the aid of algorithms(20). In addition, Yannick et al.(21) also propose that even models trained on single-center data can be robust when models need to be applied to unseen multi-center data, which provides us with non-directly against Doppelganger effect provides guidance.

- Even though there are already considerations about alignment problem, feature selection, and improving model robustness in the fields of biomedicine and artificial intelligence, current research is still committed to thinking about specific methods for specific problems but has failed to propose a general solution to Doppelganger effect. And this will be one of the problems that need to be solved in the future development of AI.

6 Conclusion

- In this report, we propose that Doppelganger effect that exists in many biomedical machine learning models and demonstrates the widespread existence of the Doppelganger effect in the entire field of data science and machine learning due to dataset shifts. It also explains how the Doppelganger effect will confound the machine learning model so that it can get inflated results in the validation set. We also introduce PPCC analysis method and propose feasible method of Doppelganger effect. To reduce the impact caused by the Doppelganger effect, this report discusses the traditional data science method, and proposes different application methods is needed to mitigate the Doppelganger effect under different problems, and believes that proper feature selection methods, which extract effective and robust data, can hedge the negative effect caused by the Doppelganger effect. We eventually come to review several existing methods and propose the limitations of feature selection problems in the biomedical domain. Future endeavor is continuously needed to tackle alignment problem.

References

[1] L. R. Wang, L. Wong, and W. W. B. Goh, “How doppelgänger effects in biomedical data confound machine learning,” *Drug Discovery Today*, 2021.

[2] R. F. Doolittle, “Convergent evolution: the need to be explicit,” *Trends in biochemical sciences*, vol. 19, no. 1, pp. 15–18, 1994.

[3] M. A. Hall, “Correlation-based feature selection for machine learning,” Ph.D. dissertation, The University of Waikato, 1999.

[4] X. Ying, “An overview of overfitting and its solutions,” in *Journal of physics: Conference series*, vol. 1168, no. 2. IOP Publishing, 2019, p. 022022.

[5] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” *arXiv preprint arXiv:1610.02136*, 2016.

[6] S. Mehrotra and V. Goyal, “Repetitive sequences in plant nuclear dna: types, distribution, evolution and function,” *Genomics, proteomics & bioinformatics*, vol. 12, no. 4, pp. 164–171, 2014.

[7] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2008.

[8] T. S. Guzella and W. M. Caminhas, “A review of machine learning approaches to spam filtering,” *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 206–10 222, 2009.

[9] S. Kumar and S. K. Singh, “Biometric recognition for pet animal,” *Journal of Software Engineering and Applications*, vol. 2014, 2014.

[10] B. Yurdakul, *Statistical properties of population stability index*. Western Michigan University, 2018.

[11] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, “Data preprocessing and intelligent data analysis,” *Intelligent data analysis*, vol. 1, no. 1, pp. 3–23, 1997.

[12] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[13] T. G. Dietterich *et al.*, "Ensemble learning," *The handbook of brain theory and neural networks*, vol. 2, no. 1, pp. 110–125, 2002.

[14] S. Singhal and M. Jena, "A study on weka tool for data preprocessing, classification and clustering," *International Journal of Innovative technology and exploring engineering (IJItee)*, vol. 2, no. 6, pp. 250–253, 2013.

[15] W. Wang, Z. Zhao, H. Zhang, Z. Wang, and F. Su, "Maskout: a data augmentation method for masked face recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1450–1455.

[16] D. van Tilborg, A. Alenicheva, and F. Grisoni, "Exposing the limitations of molecular machine learning with activity cliffs." 2022.

[17] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved problems in ml safety," *arXiv preprint arXiv:2109.13916*, 2021.

[18] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[19] S. Nagpal, S. Arora, S. Dey *et al.*, "Feature selection using gravitational search algorithm for biomedical data," *Procedia Computer Science*, vol. 115, pp. 258–265, 2017.

[20] E. Pashaei and E. Pashaei, "An efficient binary chimp optimization algorithm for feature selection in biomedical data classification," *Neural Computing and Applications*, vol. 34, no. 8, pp. 6427–6451, 2022.

[21] Y. Suter, U. Knecht, M. Alão, W. Valenzuela, E. Hewer, P. Schucht, R. Wiest, and M. Reyes, "Radiomics for glioblastoma survival analysis in pre-operative mri: exploring feature robustness, class boundaries, and machine learning techniques," *Cancer Imaging*, vol. 20, no. 1, pp. 1–13, 2020.