Doppelgänger Effects in Different Machine Learning Tasks

Guo Peihong

Abstract

Machine learning models have been increasingly used in biomedical fields. Whether it is drug development or computer aided diagnosis (CAD), machine learning provides great help for doctors and researchers. Doctors need to use high-precision CAD systems to detect possible lesions in different parts of the body for their patients, which may help patients determine the disease in the early stage, so as to formulate an effective follow-up treatment plan. Therefore, this kind of work has high requirements for the accuracy and generalization of machine learning model. However, when the data in the training set and the verification set are very similar, or the test data does not have obvious characteristic differences from the original training data, data doppelgängers will appear, resulting in good performance of the model no matter how it is trained. Some examples of doppelgänger effects in biomedical data have been presented in [1]. From the perspective of CAD in biomedicine, I analyze different kinds of doppelgänger effects, and extend them to the causes and avoidance methods of doppelgänger effects in other fields in the real world.

CAD Cases

1. Analysis

Pulmonary ground-glass opacity detection is a hot issue in biomedical detection. Data doppelgängers are ubiquitous in biomedical data, which means derived data are very similar to each other. Adenocarcinoma (ADC) [2] is the most common histological subtype of lung cancer in most countries, accounting for almost half of all lung cancers. Up to 40% of them may relapse later. Such a high chance of recurrence suggests the need to improve prognostic markers to identify early ADC patients with high-grade. Ground glass opacity (GGO) found by computed tomography (CT) is the most common lesion. GGOs may not always be obvious on CT images, and they may be omitted. The identification of GGO is based on the subjective evaluation of lung attenuation during CT, but the observation of pulmonary nodules by doctors is labor-intensive and time-consuming. Due to individual differences, the examination results are often different. Therefore, under the trend of "precision medicine", scientists use the computing model composed of multiple processing layers to learn the data representation with multiple abstract levels through deep learning. The new datasets are used by many scientists to train and test the enhanced U-NET network. Numerous experiments have shown that the approach is beneficial in improving lung nodule segmentation accuracy. This is a fantastic piece of work that is both theoretical and practical. Nowadays, in the research on GGOS, scientists use the lung image database consortium and image database resource Initiative (LIDC-IDRI) database as the source of pulmonary nodule data. This data set contains special annotation information for all nodules in complete lung CT image slices and 1007 patient image slices. However, the data at this stage is still not very comprehensive, and there is no obvious difference between different images. For example, when training a detection model of ground glass nodules, doctors mark the data and make the ground truth according to their understanding of ground glass nodules. During training and verification, the images in the datasets may not be comprehensive enough and may only include various pulmonary nodules, ignoring that the shape, size and color of other parts of the lung may also be similar to ground glass nodules. This may lead to errors or limitations in the machine learning model. The presence of data doppelgängers in both training and validation datasets inflates ML performance, even if the features are randomly selected. In addition, other nodules with different structures are still the ground glass nodules we need to detect, but we ignore them. This will result in only nodules with the same structure in the data set, resulting in model errors. Such a model is likely to perform poorly in a dataset of pulmonary nodules with similar properties but different structures.

Tumor purity estimation is crucial for accurate pathologic evaluation and for sample selection to minimize normal cell contamination in high throughput genomic analysis. Oner et al [3]. created a deep multiple instance learning model for predicting tumor purity from digital histopathology slides stained with H&E. In eight separate TCGA cohorts, their model accurately predicted tumor purity from slides of fresh-frozen sections, as well as formalin-fixed paraffin-embedded sections in a local Singapore cohort. This MIL model was developed in this study to predict tumor purity from H&E-stained histopathology sections. Each sample was represented by a bag of patches made from the sample's top and bottom slides, with the genomic tumor purity of the sample serving as the bag label. At the bag level, tumor purity was estimated, and at the sample level, representation was conducted. Although multiple types of samples (BRCA, GBM, KIRC) are included in TCGA cohorts, this does not mean that models trained on such datasets are immune to doppelgänger effects; samples with different structural but similar functional and expressive qualities may exist in the same cohort. If these cellular tissues are not taken into account, the trained model will not be generalizable and will not be able to accurately predict multiple regional samples. Furthermore, there may be cells with similar structure but different functions in different cohorts. Although nodules in the lung and certain nodules in the thyroid have a similar structure, their properties are vastly different, and their impact on the patient's health varies dramatically. If these variables aren't clearly separated, the model is likely to treat them as serious symptoms, even if they're caused by a less critical ailment. As a result, it appears that avoiding certain situations is critical.

2. Ways to Avoid

We can try to avoid these data doppelgängers in the CAD industry in two different methods.

By ensuring the diversity of datasets, the model may be trained on a wide range of samples. As stated in [1,] while creating a training dataset, try to choose data from several categories to ensure that the data are not extremely similar, resulting in better model generalization (able to have accurate predictions for different conditions). Today, ambient external elements influence the dataset preparation, and it is difficult to guarantee that the dataset comes from different categories and contains a large amount of data at the same time. As a result, there is a "valid" state when the data comes from one category, which means that the dataset can produce good training results, but there is no guarantee that there will be no doppelgänger effects, so we must try to ensure individual differences when dealing with data from the same category. Each data set, for example, originates from a distinct patient. Additionally, ensure that the degree of correlation between the training and validation sets is as high as feasible to avoid having data that is too similar. This will minimize the possibility of data doppelgängers.

Careful and precise annotation of the images ensures that Ground Truth is sufficiently representative. Samples with structural traits that are very close to those of the detection target should be given special attention, as they are likely to be otherwise irrelevant samples. This necessitates not just meticulous labeling of Ground Truth, but also an expansion of the dimensions upon which the labeling process is based. Adding various discriminators to the annotation system, in addition to structure, transparency, and size, makes the system extensive and standardized. This phase will also significantly improve the model's breadth and precision.

Optical Flow Prediction Case

1. Analysis

In recent years, optical flow estimation has advanced considerably. It finds the correspondence that occurs between the previous frame and the present frame using the change of pixels in the image sequence in the time domain and the correlation between adjacent frames, allowing the motion information of the object between adjacent frames to be determined. In addition to the Lucas-Kanade Algorithm, significant achievements for optical flow estimate have been made utilizing deep learning. Dosovitskiy et al. proposed FlowNet in [4] to use end-to-end network structure for optical flow prediction. When training the network, people mostly use datasets such as KITTI, Flying Chirs, etc. However, this is likely to cause doppelgänger effects if the representativeness of the dataset used is not excellent enough, e.g., the feature distribution of the dataset used for training is roughly within an interval, while the image features of the validation and training sets do not differ much. I tried to train FlowNet and predict the results in my prior optical flow estimating task. Because all of my training sets had the same features and a single color background. As a result, my model was not generalizable. When creating the test set, I chose images that were similar to the training set, which resulted in outstanding test results and led me to conclude that the model had been properly trained, as shown in the first two columns of Figure **1**. However, when I used images with slightly complex background, the prediction of the model deviated greatly, as indicated in the third column of Figure **1**.

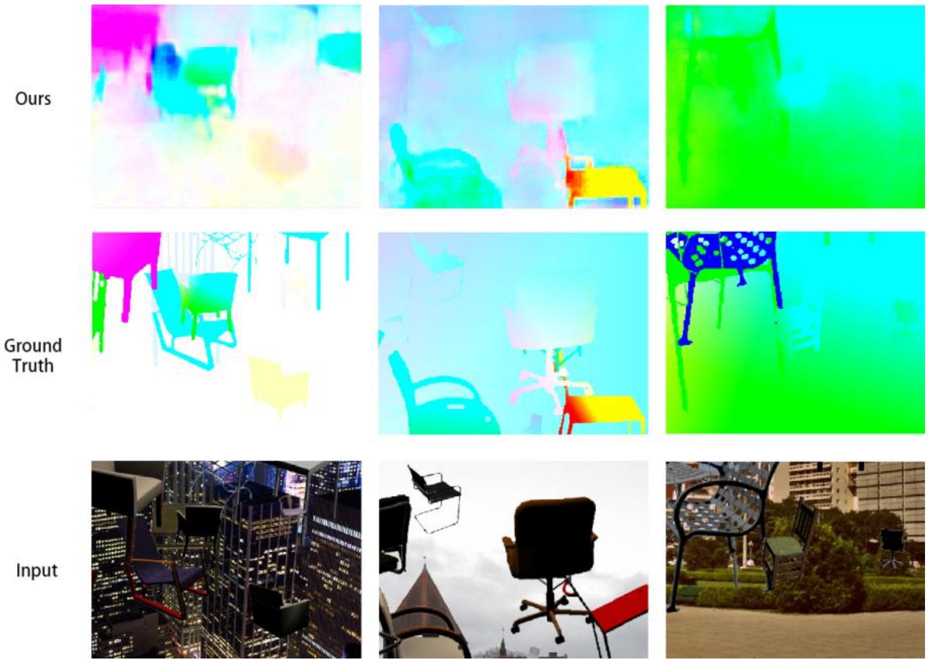

Fig 1. Visualization of flying chairs test results

2. Ways to Avoid

The training set should include images of different complexity. Only after learning a wide variety of features, the model can show robustness.

We should Improve the structure of deep neural networks, especially the performance of encoder. FlowNetCorr in [4] provides us with a good example. We need to innovate the network structure so that the network can learn very small features, and it is these small features that lead to differences across data. The network structure is improved to better extract the image features in the datasets by dimensionality reduction.

Conclusion

In this paper, I discuss in detail the doppelgänger effects in today's biomedical data from Pulmonary ground-glass opacity detection and Tumor purity estimation in CAD. Regarding the reasons for the appearance of doppelgänger effects, I summarize them as **1**. the limitations of biomedical data, **2**. the training samples are not extensive and omissions or redundancies occur in the selection of samples. **3**. the high correlation between the training and validation sets.

Besides biomedicine, I also analyze data doppelgänger in the optical flow prediction task and conclude that doppelganger effects are not unique to biomedical data.

Taking all the analyses together, I propose specific avoidance options: **1**. By ensuring the diversity of datasets, the model may be trained on a wide range of samples. **2**. Carefully and accurately annotate the image. **3**. Improve the structure of deep neural networks, especially the performance of encoder.

References

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

[2] Wang X, Zhang L, Yang X, et al. Deep learning combined with radiomics may optimize the prediction in differentiating high-grade lung adenocarcinomas in ground glass opacity lesions on CT scans[J]. European journal of radiology, 2020, 129: 109150.

[3] Oner M U, Chen J, Revkov E, et al. Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study[J]. bioRxiv, 2021.

[4] Fischer P, Dosovitskiy A, Ilg E, et al. Flownet: Learning optical flow with convolutional networks[J]. arXiv preprint arXiv:1504.06852, 2015.