Report 3

The Doppelganger effects is not unique to biomedical data. In the presented paper, the author proposes that the doppelganger phenomenon occurs in the process of model validation because the training set and validation set are not independent, which is often assumed to be. We need to consider whether the data is offset, that is, the data distribution of training set and verification set is different.

Covariate Shift is a common type of dataset shift. Covariates are the input variables (independent variables) of the model. Covariate Shift refers to the fact that input variables in the training set and validation set have different data distributions, meaning that only the input distribution changes while the input-to-output mapping remains the same. Ideally, the training set and the verification set should have the same data distribution, but in reality, many external factors may lead to this assumption is not true.

In image data, factors affecting data distribution may include:
1. **Category ratio**: For example, 30% cars, 40% people and 30% trees in the training set and 10% cars, 20% people and 70% trees in the test set, we can guess that the data in the training set might come from street view and the data in the test set might come from the park. The distribution of data is different because the proportions of the two classes are different.

Figure1. Different Category Ratios

2. **Image features**: For example, the training set contains very clear cat and dog images collected by us on the Internet, while the test set contains vague cat and dog images taken by ourselves; Or if the training set is an image of a white dog and a yellow cat, and the test set is an image of a yellow dog and a white cat, then they can also be viewed as having different data distributions.

Assume a 96:2 split for the training/verification set:

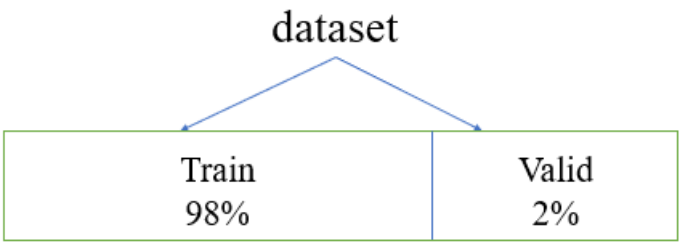

Figure 2. Training set and verification set

Overfitting and data mismatch

Assume a training error of 2% and a verification error of 10%. Given that the two sets of data come from different distributions, it is not possible to judge how much of the 8% difference between the two sets of data is due to data mismatches and how much is due to model overfitting.

To solve this problem, let's modify the training/verification/test set partition. Take out a small part of the training set and call it the bridge set. This part will not be used to train the network model. It is an independent set.

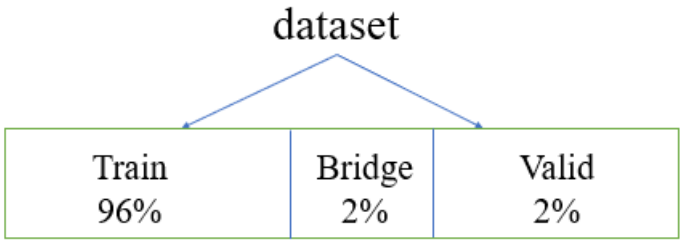

Figure 3. Bridge Set

Overfitting

In this way, if the training error and verification error are 2% and 10% respectively, while the bridge set error is 9%, because the bridge set and the training set come from the same distribution, excluding the influence of data mismatch (data distribution), the error difference between them is 7%, so 7% of the error comes from the variance error (overfitting). There is a 1% difference between the bridge set and the validation set, so 1% of the error comes from data mismatch errors.

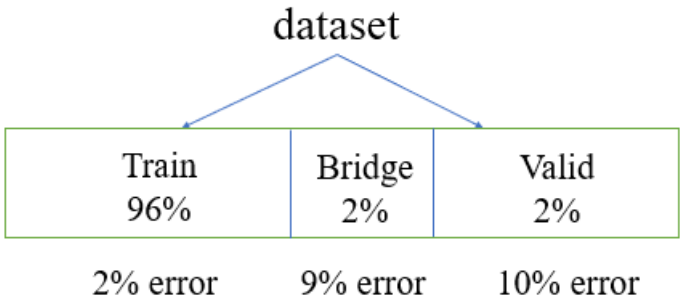

Figure 4. Error in the First Case

Among the 8% errors between the training set and the validation set, 7% are variance errors and 1% are data mismatching errors.

Data Mismatch

Now assume that the bridge set error is 3% and the rest is shown above:

According to the above analysis, the error difference between the training set and the bridge set is 1%, so 1% of the error comes from the variance error. There is a 7% difference between the bridge set and the validation set, so 7% of the error comes from data mismatch errors.

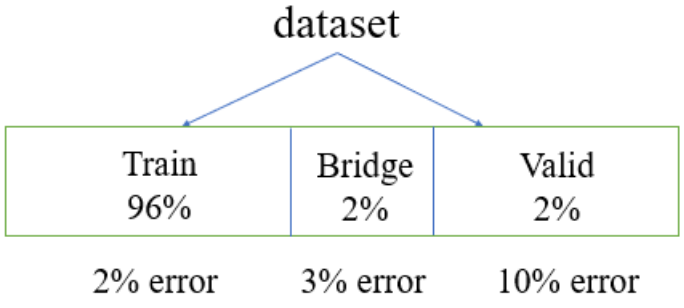

Figure 5. Error in the Second Case

Reducing variance error is a common task in machine learning. For example, various regularization methods can be used.

Solutions

For data mismatching errors, we can try to collect more real data and add them to the training set to participate in model training, or generate data very similar to real data as much as possible and add them to the training set.