Propose of Doppelganger Mining Method

Doppelganger effects will be highly likelihood to occur when there are similarities in training and test data set. In biomedical data science, this similarity usually emerges from a seemingly correct but not yet proved assumption. For instance, proteins of similar structures will always have analogous functions, whereas in some cases those proteins that are quite 'unsimilar' in shape or inherited from different ancestors can also play the same role in cells. Subsequently, this will falsify the true effectiveness of machine learning models in biomedical data science and inflation takes place.

I want to elaborate on this topic from two perspectives: the presence of this doppelganger effect in other machine learning applications and one potential method to mitigate its negative impact over model effectiveness.

It is nothing more than a naive assertion that this doppelganger effect will be encountered in face recognition scenarios using deep convolutional neural networks as it is based on such an assumption that if two matrices, individually representing two images, have a close cosine distance, then it is probable that these two matrices depict the two same faces. This assumption is true in most cases, but bad situations will always be encountered when two people share very similar embedding faces when there is a low-shot or etc. This misclassification can be dangerous when face recognition is for online payment, personal info censor etc. I say this is a naive assertion because that is the logic of training a neural network, similar feature maps will always output a similar predictions value. In face recognition applications, there is large-scale data available and more spatially variated representation of the same face accessible, and this data convenience can bring benefits to train model with more sophisticated parameters and thereby improve the accuracy and correctness of the model. In other words, a model in face recognition can readily have more data and more knowledge.

In my first few touches of biomedical data science and its publications, I fell in love with this subject as it required more technologies to design an efficient model when data is sometimes controversial and not available for researchers. Please pardon my rudeness if it is not true when data collection also requires much expertise.

In addition to feeding more data to our machine learning model, there are some methods investigated to mitigate the impact of Doppelganger effect. One of them I want to reference in this report and may be further referenced in biomedical data science is Doppelganger Mining [1].

Doppelganger-Mining is a strategy applied to track doppelganger identities in our training data set. In other words, we not only treat all data input as individual samples, but also care about the correlation among them which corresponds to data similarities. Since this is a face recognition sampling method, a non-static list is created to keep record of most identical identities of the sample data. In biomedical data science, this method can be reused to record potential doppelgangers and put forward related strategies for these doppelgangers. But how can we identify identities or data samples that are most analogous?

At the training setup process, exemplar-based loss functions will be beneficial in determining the relationship among samples. Since these examples are randomly distributed among training, validation and test sets, it is indispensable to determine this relationship prior to training process or apply cross-validation otherwise to fully figure out similarities among samples. An exemplar-based loss function tries to use example-to-example distance information to train the networks and the distance mentioned is usually measured by a loss function. There are a couple of feasible loss functions available in face recognition applications, but its usage in biomedical data science field requires more expertise. It will be my honor should there be an opportunity to resume investigation into mitigating the doppelganger effect in health and medical machine learning projects with appropriate loss functions despite them being not accessible now.

Given the correlation among individual samples, now it is time to resume training process. A very intuitive way of avoiding doppelganger effects is to place those analogous samples into the same set instead of two. Other methods are also reserved for further discussion.

In addition, this method requires extra computation time and computation resources which seems to be less efficient than traditional training methods. Hence, another obvious optimization is to train 'correlation' parameters simultaneously when training the original neural network. This is an interesting topic but more complicated to implement from my perspective.

References

1. Smirnov, E., Melnikov, A., Novoselov, S., Luckyanets, E., & Lavrentyeva, G. (2017). Doppelganger mining for face representation learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 1916-1923).