知の物理学研究センター

ipi seminar

Institute for Physics of Intelligence

Ziyin Liu

（東京大学大学院理学系研究科物理学専攻/
The University of Tokyo, Department of Physics）

「Careful Deep Learning: Learning to Abstain by Training on A Simple Loss Function」

２０１９年10月17日(木)　１０時３０分～１２時００分

東京大学本郷キャンパス理学部１号館９階 913 号室

Deep learning achieved remarkable success in classification problems. However, the current training methods for classification forces neural networks to predict a class on whatever input it receives.

However, there exists cases when a prediction should not be given. For example, a neural network trained to predict digits should refrain from prediction when seeing a picture of an airplane. We transform the original m-class classification problem to (m+1)-class where the (m+1)-th class represents the model abstaining from making a prediction due to uncertainty. Inspired by portfolio theory, we bridge a connection between gambling and making prediction. We propose a loss function to solve this problem based on the doubling rate of gambling. We show that minimizing this loss function has a natural interpretation as maximizing the return of a horse race, where a player aims to balance between betting on an outcome (i.e., making a prediction) when confident and reserving one's winnings (i.e., abstaining) when not confident. This loss function allows us to train neural networks and characterize the confidence in prediction in an end-to-end fashion and achieve very competitive result.