Applying influence function to various neural network models

Keun Hoi Ahn[1], and Myungjoo Kang[2]

1) *1) Department of Computational Science and Technology, Seoul National University, Seoul, Republic of Korea*
2) *Department of Mathematical Sciences, Seoul National University, Seoul, Republic of Korea*

Corresponding Author : Myungjoo Kang, mkang@snu.ac.kr

ABSTRACT

Most of Black-box models often work very well but we still cannot develop reasoning method for them. To apply real world through various AI service, researchers work very hard to explain the mechanism of deep learning models' black-box prediction. In this poster, we try to adopt influence function to loss functions of various models to show the effect of influence function.

REFERENCES

1. Pang Wei Koh and Percy Liang., "Understanding Black-box Predictions via Influence Functions," *International Conference on Machine Learning, 2017*.