



Contribution ID: 44

Type: **not specified**

Damien Garreau - What does LIME really see in images?

Friday, October 7, 2022 10:00 AM (1 hour)

The performance of modern algorithms on certain computer vision tasks such as object recognition is now close to that of humans. This success was achieved at the price of complicated architectures depending on millions of parameters and it has become quite challenging to understand how particular predictions are made. Interpretability methods propose to give us this understanding. In this paper, we study LIME, perhaps one of the most popular. On the theoretical side, we show that when the number of generated examples is large, LIME explanations are concentrated around a limit explanation for which we give an explicit expression. We further this study for elementary shape detectors and linear models. As a consequence of this analysis, we uncover a connection between LIME and integrated gradients, another explanation method. More precisely, the LIME explanations are similar to the sum of integrated gradients over the superpixels used in the preprocessing step of LIME.