

ПРОБЛЕМИ РИЗИКУ ТА БЕЗПЕКИ В ГЛИБОКИХ МЕРЕЖАХ



HOME LECTURES ABOUT US SIGN IN SIGN UP

Risk and Safety Concerns in Deep Networks - ADAMS WAI-KIN KONG

In recent years, deep networks have demonstrated their powerful capabilities, often outperforming human experts, such as AlphaGO. Many have been deployed to real applications, creating great business value. Because of their great impact to human beings and our society, many have concerned about their potential risks. In this talk, the speaker will first give some background information about adversarial attacks, including black-box and white-box attacks and attribution methods, including integrated gradients. Then, he will discuss how to use attribution methods, in particular integrated gradients, which was originally developed for explaining deep methods, to study risks of transferable attacks. Similar to deep network output, attributions can be manipulated by adversarial attacks. In the second half of this talk, the speaker will discuss how to protect attributions against adversarial attacks



Adams Wai-Kin Kong
Associate Professor, director of Master of Science in Artificial Intelligence programme, Nanyang Technological University, Singapore

Register for free!

6 травня 2024 року відбудеться онлайн-лекція «Проблеми ризику та безпеки в глибоких мережах» директора магістерської програми зі штучного інтелекту Наньянського технологічного університету.

Планується розглянути наступні питання: змагальні атаки, включно з атаками «чорної скриньки» та «білої скриньки»; методи атрибуції та інтегровані градієнти; способи використання методів атрибуції для пояснення глибинних методів і вивчення ризиків перенесених атак; виявлення фактів маніпуляції за допомогою агресивних атак; інструменти захисту. Організатор – Ukraine Global Faculty.

Детальніше: <http://surl.li/skisin>, <http://surl.li/sxoow>

Фото: UGF

#НРАТ_НауковіЗаходи

6 травня 2024 року відбудеться онлайн-лекція «Проблеми ризику та безпеки в глибоких мережах» директора магістерської програми зі штучного

інтелекту Наньянського технологічного університету.

Планується розглянути наступні питання: змагальні атаки, включно з атаками «чорної скриньки» та «білої скриньки»; методи атрибуції та інтегровані градієнти; способи використання методів атрибуції для пояснення глибинних методів і вивчення ризиків перенесених атак; виявлення фактів маніпуляції за допомогою агресивних атак; інструменти захисту. Організатор – Ukraine Global Faculty.

Детальніше: <http://surl.li/skIsn>, <http://surl.li/sxoow>

Фото: UGF

#НРАТ_НауковіЗаходи

2024-05-03

Інформація з офіційного вебпорталу Національного репозитарію академічних текстів