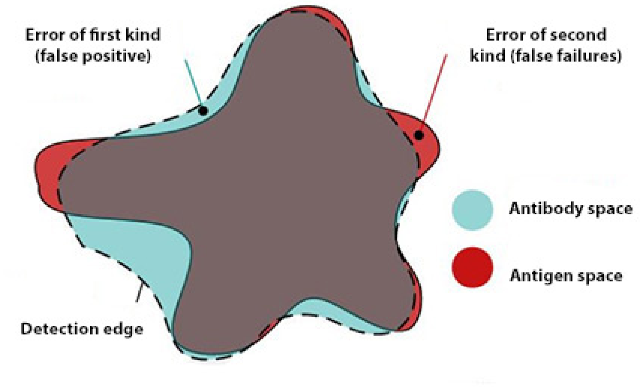
Digital immunity: SibFU scientists to teach AI to catch data leakage

SibFU scientists are developing a self-learning artificial intelligence system capable of detecting hidden information inbuilt in various media content disseminated in Internet resources. This artificial immune system mimics some of the natural processes that take place in the human immune system, for instance, it produces special digital antibodies to detect hostile antigens — media files which covertly embed additional information. The main results of the study are given in the article here.


Error of first kind (false positive)
Error of second kind (false failures)
Antibody space
Antigen space
Detection edge

> *‘Steganography is a very ancient art of cryptography. The shaving of a secret message on the slave’s head, sympathetic ink that becomes visible only in ultraviolet — all these are the first attempts of mankind to make a double-bottomed message, unobvious to the untrained eye. With the development of the Internet and various social services since the mid-2000s, truly gigantic volumes of media data have begun to circulate in the network, including images in JPEG — one of the most popular formats. In our research, we focused on the methods of discovering what is hidden behind a simple, at first glance, picture,’* said **Aleksey Shniperov**, assistant professor of the Department of applied mathematics and computer security, Siberian Federal University.

The scientist explained that it is impossible to distinguish a clean image from that one filled with hidden information without special tools — only knowing the way any additional content has been built in will make it possible to extract, for example, a map of a strategic object. Or detect a hidden extremist message. For a layman, an image of a tree or a smiling girl, fit out with a secret stuffing, will not be any different from millions of such files.

> *‘Nowadays, the problem of sending potentially dangerous media content is gaining momentum and troubles, among other things, large companies and corporations. Lest become victims of information leaks and to avoid large financial losses, they spend formidable amounts of money on developing systems that can filter the content of the World Wide Web. The complexity of working with media files lies in their diversity. Those same very JPEG images have got a huge number of parameters: their content, colour rendering parameters, compression ratio, etc. Moreover, over the past decades, the number of ways to embed secret content into a usual picture, including ready-made software, has been growing like an avalanche. So far, there is no single superalgorithm that, like Superman, will instantly save the network from double-bottom content. But here at SibFU, we are developing one of the possible tools to combat such images. The system is built similarly to the human immune system, which has proved its efficacy, and therefore is able to detect files which seem suspicious for it, much as our blood cells detect a virus or a bacterium,’* the scientist continued.

The immune network created by the Krasnoyarsk researchers, similar to human immunity, does not know for sure which breacher it will have to face. It does not get stuck on useless signs such as size or an object shown in the picture (they will be absolutely identical in the safe and hazardous files), but it constantly reproduces a kind of antibodies which differ from each other so that it will be more likely to detect unsound files which may hide additional content.

> *‘The developed system is constantly self-learning, producing non-stop new types of digital antibodies, each of which is a multidimensional space of specific vectors-parameters. When one of them suddenly clings on the off-chance of some kind of media file, we, as digital doctors, check what is wrong with this file. If this antibody worked correctly and managed to detect hidden content in a JPEG file, for example, then digital antibodies similar to it will be produced by the system in larger quantities. In this case, our network considers images not loaded with hidden content as normal objects. Digital immunity simply neglects them,’* told **Aleksey Shniperov**.

Another similarity to biological immunity can be found in the jabs that developers give to the system under development: they load a certain number (for example, several hundred) of unsound files into it, and this helps the system develop acquired digital immunity, much like a flu vaccine mobilizes the body's defences and protects against an outbreak of seasonal diseases.

> *‘A selection of dangerous images which we use to train our artificial immune system is a drop in the sea of media files circulating in the Net. Of course, the system learns quickly as digital antibodies multiply, and their negative and clonic selection is carried out — the system rejects those ones that can no longer be considered antibodies by a number of signs and are not suitable for detecting an enemy. But now we have reached the limit of the technical capabilities of our equipment: more powerful chipsets are required for digital immunity to train further. The next step will be the transfer of all calculations to thousands of specialized processing units that are available in any modern graphics adapter or their clusters,’* concluded the researcher.

By the way, another problem of digital immunity may be the analogue of an autoimmune reaction — digital antibodies are still imperfect and can sometimes still respond to harmless images. Such digital allergy, according to the SibFU scientists, can develop with a significant increase in working antibodies in the artificial immune system after moving to more up-to-date computers, which will require additional research and adjustments to the model of the digital immune system.

“Scientific Russia” wrote about the formation of digital immunity by the scientists of Siberian Federal University.

11 february 2020

Web page address: https://news.sfu-kras.ru/node/22765