Deep Generative View on Continual Learning

What if we had no memory?

In 1953, a patient named Henry Molaison suffering from a debilitating epilepsy was admitted by doctors for a risky surgical procedure. In an attempt to remove the cause of his illness, the surgeons literary sucked out part of Henry's brain. The procedure that aimed at helping Henry turned out to be a disaster for him. Even though the patient no longer suffered from epilepsy, he permanently lost his ability to memorize new experiences. He could still remember some memories – scenes from his childhood, facts about his parents, and historical events – but only if they had occurred before the surgery.

Henry lived the most peculiar life one can imagine. He remembered only very distant memories while waking up everyday as if he was just born. He was able to improve his performance on motor tasks, even though he had no recollection of ever practicing them. During all this time, he was the subject of extensive studies, and his traumatic experience led to one of the most significant discoveries in contemporary neuroscience.

What if I told you, that even the most advanced artificial intelligence systems suffer from the very same symptoms as Henry Molaison?

Artificial memories

A machine that could think like a person has been the Holy Grail of AI research since its earliest days. A tempting vision of a real-life commander Data from Star Trek or C3PO from Star Wars is, however, not possible yet. This is largely because of several attributes of human intelligence that recent artificial systems still miss.

In principle, because of the lack of properly simulated artificial memory, intelligent systems currently in use are typically based on the assumption that once the model is trained to perform a given task, it should be able to execute it indefinitely, without the need for any further training. While this assumption simplifies the problem of training, it also distances us further away from the ultimate goal of artificial intelligence research.

Moreover, it is even more problematic when taking into account the practical aspect of machine learning systems. The ability to learn and improve from new observations without forgetting the previous ones is critically needed in many domains such as *e.g.* in autonomous driving. Contemporary methods including neural networks are very capable of learning new skills. However, they achieve that while catastrophically forgetting previously acquired knowledge at the same time.

Would you really get into the car with Henry Molaison, who can't even remember if he ever learned how to drive?

Three ways to remember

In this project, we plan to focus on the lack of proper memory storing and consolidation mechanisms in contemporary machine learning models. Recent works indicate three possibilities of how we can avoid forgetting. The first one, known as regularization, tries to slow down the process of training of the most crucial parts of the artificial brain, not to overwrite what is already stored in there. This approach, however, only slows down the process of forgetting and does not really prevent it. The second one expands artificial neural networks with each new trained task. While methods based on this idea yield good results, this is not a valid method when considering lifespans of systems we want to continually train. Finally, the third method is based on the assumption that the best way to remember previous knowledge is to rehearse it from time to time. Nevertheless, storing a significant number of previously seen examples may consume a tremendous amount of space.

In this project, we once more benefit from the experience of Henry Molaison, by drawing inspiration from the neuroscience discoveries triggered by his condition. Therefore, we propose to create a new, general solution which will be able to incrementally store new memories consolidating them with the previous ones instead of overwriting them. To that end, we want to develop new neural networks based on generative models that will be able to incorporate and consolidate knowledge in the same way humans do.