Nr rejestracyjny: 2023/49/B/ST6/02580; Kierownik projektu: dr hab. inż. Paweł Kacper Morawiecki

Ochrona danych i prywatność a generatywne sieci neuronowe

W ostatnich latach nastąpił szybki postęp w technikach modelowania generatywnego w dziedzinie głębokiego uczenia. Wśród nich generatywne modele dyfuzyjne szczególnie zyskały na znaczeniu ze względu na ich zdolność do generowania wysokiej jakości, różnorodnych i skomplikowanych obrazów. Modele te już są szeroko stosowane w marketingu, wspomaganiu artystów, grafice komputerowych, projektowaniu gier. Jednak w miarę jak modele te stają się coraz powszechniej stosowane, pojawiają się niepokojące kwestie prywatności i własności danych. Przykładowo w tym roku firma Getty Images złożyła pozew przeciwko Stability AI, oskarżając ją o nielegalne kopiowanie i przetwarzanie milionów obrazów chronionych prawem autorskim.

Jednym z krytycznych problemów pojawiających się w tym kontekście jest ustalenie, czy podczas procesu uczenia modelu wykorzystano określoną grafikę czy nie. Wydobycie tych informacji z modelu może mieć kluczowe znaczenie w przypadkach, gdy dane chronione prawem autorskim lub dane wrażliwe są wykorzystywane bez pozwolenia. Wagę tych kwestii odzwierciedla Ogólne Rozporządzenie Unii Europejskiej o Ochronie Danych, w szczególności artykuł 17 często określany jako „prawo do bycia zapomnianym".

Obecnie najlepsze modele generatywne są warunkowane tekstowo, a mianowicie użytkownik dostarcza tekst, np. „obraz statku kosmicznego w stylu Van Gogha" i model generuje obraz. Modele te z łatwością naśladują styl danego artysty, ale są też na tyle elastyczne, że pozwalają na jego modyfikację lub umieszczenie obrazu w niecodziennym kontekście.

Figure 1: Nowoczesne duże modele generatywne bardzo rzadko ślepo kopiują obrazy ze zbioru uczącego. Zamiast tego elastycznie wyodrębniają koncepcje, takie jak styl, ton lub tekstura, aby generować nowe obrazy. Oryginalne dzieło Zdzisława Beksińskiego (po lewej) vs obraz wygenerowany przez model generatywny Midjourney tekstem „Beksinski vision of love" (po prawej). Wygenerowany obraz zachowuje charakterystyczne cechy sztuki Beksińskiego, a mianowicie specyficzną fakturę i wszechobecny rozkład.

Prowadzi to do sytuacji, w której niezwykle trudno jest wywnioskować, czy dany obraz został wykorzystany w treningu czy też nie. Dlatego trudno jest ustalić własność danych i zaczynają pojawiać się obawy dotyczące prywatności.

W tym projekcie chcemy zbadać, czy możliwe jest wywnioskowanie znaczących informacji na temat zbioru uczącego dla dużych generatywnych sieci neuronowych. Interesuje nas również oduczenie (zapomnienie) danego obrazu, tak aby sieć neuronowa nie była już w stanie wygenerować bardzo podobnego obrazu lub jego stylu. Takie precyzyjne zapomnienie, gdyby się powiodło, pozwoliłoby na zapomnienie/wycofanie prac użytkownika bez konieczności ponownego, bardzo kosztownego uczenia całej sieci.