Szalona SI

Modele sztucznej inteligencji bazują na treningu wykorzystującym ogromne zbiory danych. Co się jednak stanie, jeśli te dane będą coraz niższej jakości? Czy SI czeka informacyjne zatrucie?

W połowie lat 80. ubiegłego wieku stada krów w Wielkiej Brytanii zaczęła dotykać tajemnicza choroba. Zwierzęta zachowywały się dziwnie – niektóre stały się agresywne, inne wpadały w otępienie. Wykryto u nich również zmiany neurologiczne. Tę nową przypadłość media szybko ochrzciły mianem „choroby szalonych krów", a jej naukowa nazwa to gąbczasta encefalopatia bydła, czyli BSE (Bovine Spongiform Encephalopathy).

Badacze dość szybko wytropili, że chore zwierzęta karmione były białkowymi preparatami pozyskanymi z ciał innych krów. Powstała wtedy również teoria tłumacząca przenoszenie się choroby – wywoływać ją mają priony, białkowe cząsteczki zakaźne odkryte zaledwie kilka lat wcześniej. Za zidentyfikowanie prionów i wskazanie ich potencjału chorobotwórczego Stanley Prusiner otrzymał zresztą Nagrodę Nobla w dziedzinie medycyny.

Piotr Kościelniak
dziennikarz, popularyzator nauki

W ciągu kilku lat BSE rozprzestrzeniła się w całej Europie – z jej powodu wybijano całe stada bydła. Ale to dopiero początek horroru, bo bardzo podobną chorobę wykryto również u ludzi. W latach 90. XX wieku pojawił się wariant choroby Creutzfeldta-Jakoba – vCJD – gąbczaste zwyrodnienie mózgu prawdopodobnie powodowane przez priony. Skąd priony? Ze zjadanego przez ludzi mięsa zakażonych BSE krów. Ta choroba, na szczęście rzadka, rozwija się wiele miesięcy, ale zawsze prowadzi do otępienia, depresji, drżenia mięśni, halucynacji i wreszcie do śmierci.

Dlaczego w artykule o sztucznej inteligencji piszę o krowach, prionach, gąbczastym zwyrodnieniu mózgu i halucynacjach? Bo ten sam fenomen dotyka dziś modeli SI karmionych przetworzoną syntetyczną papką danych.

> *Najnowsza, najbardziej obiecująca technologia, która ma zrewolucjonizować każdą dziedzinę naszego życia, obciążona jest bowiem mechanizmem autodestrukcyjnym.*

Autofagi i Uroboros

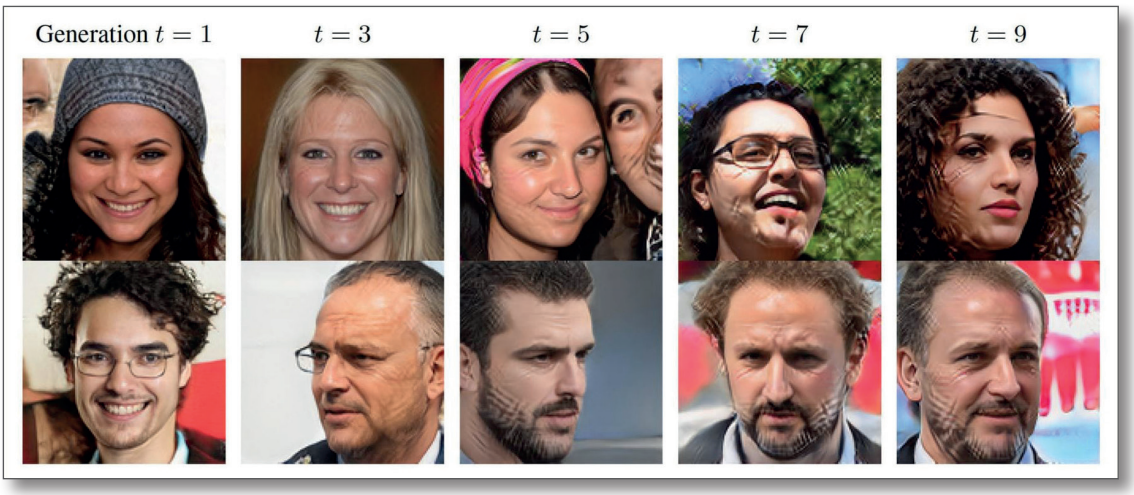
Najbardziej zaawansowane modele sztucznej inteligencji, zwłaszcza takie, które mają operować na bardzo szerokich zakresach danych – jak duże modele językowe (LLM) – karmione są ogromnymi zbiorami informacji. Informacje te są pobierane z internetu: czasem ze specjalnie przygotowanych baz danych, czasem po prostu ze stron WWW. A kiedy te dane się kończą, do treningu wykorzystane są dane syntetyczne – specjalnie przygotowane zestawy traktowane jako uzupełnienie danych rzeczywistych.

Takie dane są po pierwsze tańsze od tych, które przygotowali ludzie, a po drugie – nie obciążone prawami autorskimi. Firmom, które z nich korzystają, nie grożą zatem pozwy od np. zirytowanych wydawców gazet, których artykuły zostały bezprawnie wykorzystane do szkolenia SI. To zaś sprawia, że firmy budujące i zarabiające na sztucznej inteligencji z takich zestawów nader chętnie korzystają.

Mało tego, każda kolejna generacja modeli sztucznej inteligencji siłą rzeczy korzysta z danych, które wcześniej przygotowały poprzednie modele. W ten sposób specjaliści od SI starają się zaspokoić olbrzymi apetyt na dane sztucznej inteligencji. I tak następne pokolenia SI otrzymują materiał wejściowy o coraz niższej jakości. Stara prawda: *garbage in, garbage out,* czyli śmieci na wejściu – śmieci na wyjściu kolejny raz znajduje potwierdzenie.

Badania nad najnowszymi generacjami modeli sztucznej inteligencji wskazują, że trenowanie SI wykorzystujące dane wykreowane przez SI w krótkim czasie prowadzi do degeneracji. W jak krótkim? Analizy przeprowadzone przez naukowców z uniwersytetów Stanforda i Rice'a dowiodły, że wystarczy do tego pięć cykli tworzenia danych przez model i trenowania go na ich podstawie.

„Ogromne postępy w dziedzinie algorytmów generatywnej SI dla obrazów, tekstów i innych rodzajów treści wprowadziły pokusę wykorzystania syntetycznych danych do trenowania modeli kolejnych generacji. Powtarzanie tego procesu doprowadziło do pętli autofagii – samozjadania SI, który nie jest jeszcze dobrze poznany" – piszą autorzy analizy „Self-Consuming Generative Models Go MAD". „Nasza najważniejsza konkluzja jest taka, że bez wystarczającej ilości świeżych, prawdziwych danych wstrzykiwanych do pętli, kolejne wersje modeli generatywnej sztucznej inteligencji skazane są na produkowanie



W tym przykładzie naukowcy przeszkolili szereg modeli generatywnych StyleGAN-2 przy użyciu w pełni syntetycznych danych. Każda kolumna obrazu wyświetla kilka przykładów wygenerowanych odpowiednio przez model pierwszej, trzeciej, piątej, siódmej i dziewiątej generacji. Z każdą iteracją pętli zakreskowane artefakty są stopniowo wzmacniane (zdjęcie dzięki uprzejmości Digital Signal Processing Group/Rice University).

treści o coraz niższej jakości i coraz mniejszym zróżnicowaniu. Nazwaliśmy to zjawisko Model Autophagy Disorder (MAD, szalony – przyp. red.), nawiązując do choroby szalonych krów".

Naukowcy wybrali bardzo widowiskową metodę dowodzenia swojej tezy. Za pomocą SI wygenerowali obrazy ludzkich twarzy o fotorealistycznej jakości. Następnie dodawali te obrazy do zestawu danych służących do uczenia kolejnych generacji SI. Bez trudu dała się zauważyć generalna degradacja jakości obrazu – sztuczna inteligencja wprowadzała nowe cechy do obrazu (artefakty), które zniekształcały wizerunki ludzi – wyglądały jak blizny na twarzach.

W praktyce model uwypukla artefakty z każdą kolejną generacją. Równocześnie ucina informacje, które uznaje za najmniej istotne (znajdujące się na skrajach histogramu), a podkreśla te, które są najbardziej uśrednione. Trochę przypomina zatem mitycznego węża Uroborosa, który wiecznie pożera własny ogon.

Wieże i zające

To samo zjawisko zauważyli wcześniej naukowcy z Uniwersytetu Oksfordzkiego i opisali je w artykule „AI models collapse when trained on recursively generated data" w prestiżowym magazynie „Nature" w połowie 2024 r. Szef zespołu, Ilia Shumailov twierdzi, że efekt ten wprawdzie nie spowoduje zawalenia się dotychczasowych systemów SI, ale doprowadzi do spowolnienia postępu i znaczącego obniżenia jakości materiałów generowanych przez systemy sztucznej inteligencji.

Zespół Shumailova eksperymentował z jednym z dużych modeli językowych, który wytrenowano na zawartości Wikipedii. Następnie „udoskonalano" jego działanie, karmiąc go jego własnymi wcześniejszymi wynikami. I tak przez dziewięć generacji. Naukowcy opracowali również wskaźnik absurdalności tekstu – im głupsza i mniej zrozumiała była odpowiedź sztucznej inteligencji, tym wskaźnik wyższy.

Pytania (prompty) dotyczyły dokończenia zdań – na przykład części tekstu traktującego o architekturze wież kościołów parafialnych i pracy zespołów murarskich w XIV w. O ile pierwsza generacja radziła sobie z tym bez problemu, o tyle dziewiąta pogubiła się zupełnie. Zamiast o architekturze opowiadała o… populacjach różnych gatunków zajęcy. A czasem było nawet jeszcze gorzej – część liter zastąpiła znakami specjalnymi, co sprawiło, że odpowiadała nie na temat i w sposób niezrozumiały. W przeciwieństwie do pierwszej generacji, ta ostatnia, karmiona przetworzonymi danymi, uzyskała najwyższy wskaźnik absurdalności wyników.

„Jeśli zrobisz zdjęcie, wydrukujesz, a następnie je zeskanujesz i znowu wydrukujesz, i będziesz ten proces powtarzać wiele razy, to w końcu dostaniesz tylko szum i zakłócenia" – tłumaczy Ilia Shumailov z rozmowie w „MIT Technology Review".

Habsburgowie i ChatGPT

Jakie są przyczyny tego problemu? To przede wszystkim skala zbiorów danych potrzebnych do wytrenowania kolejnych generacji modeli.

„Efektywność funkcjonowania modeli zależy od skali zbiorów danych. To dlatego do szkolenia SI wykorzystywane są syntetyczne dane stworzone w specjalnie przygotowanym, kontrolowanym środowisku. Przekopywanie internetu w poszukiwaniu nowych danych przynosi bowiem coraz słabsze rezultaty" – opisuje problem ze szkoleniem sztucznej inteligencji Shayne Longpre z MIT Media Lab.

Jeden z pierwszych modeli tłumaczących (typu transformer) wykorzystywał dane będące równowartością treści z ok. tysiąca książek. Pierwsza wersja ChatGPT – ok. 7 tys. książek. Każda kolejna wersja tego modelu potrzebowała coraz więcej – rozmiar zestawu danych, który SI musiała „przeczytać", rósł wykładniczo. Model GPT-3 wytrenowany na 3 mld stron internetowych (zestaw Common Crawl) w 2020 r. operował na ok. 300 mln tokenów (dla uproszczenia potraktujmy token jako słowo w książce). GPT-4 jest jeszcze ponad dziesięć razy większy.

Pomijając fakt, że zapoznanie się z taką ilością informacji przez człowieka jest nierealne choćby z tego powodu, że musielibyśmy żyć kilka tysięcy lat, dostępność tak ogromnej ilości danych wysokiej jakości – oryginalnych – jest już ograniczona. Dlatego naukowcy przygotowujący modele AI sięgają po dane wytworzone przez inne modele SI.

A to prowadzi nas od krów i prionów w stronę innej analogii – dynastii Habsburgów, w której bliskie kazirodczym stosunki zawiązywane w kolejnych pokoleniach doprowadziły od uwypuklenia pewnych cech wyglądu (słynna warga habsburska) czy schorzeń genetycznych. „Habsburg SI to taki system, który tak mocno polega na danych wytworzonych przez inne generatywne systemy SI, że staje się mutantem, efektem chowu wsobnego o przerysowanych, groteskowych cechach" – napisał Jathan Sadowski, który zajmuje się m.in. społeczną teoria informacji i problematyką trenowania sztucznej inteligencji. „Nadal otwarte jest pytanie: jak wiele syntetycznych danych to już za wiele. I będą sobie musiały na to pytanie odpowiedzieć największe firmy budujące systemy SI oraz inżynierowie" – mówi Sadowski.

Kopernik i internetowy śmietnik

Dokąd nas prowadzi przetwarzanie przeżutej przez SI papki informacyjnej? Kierunek jest dość niepokojący, zwłaszcza jeśli wziąć pod uwagę to, że różne systemy SI są coraz chętniej

wykorzystywane przez wielkie firmy – począwszy od finansów, ubezpieczeń i księgowości przez medycynę i inżynierię, a skończywszy na mediach i rozrywce. Przypadłością „szalonych procesorów" są dotknięte bowiem nie tylko duże modele językowe, z którymi większość z nas ma do czynienia, lecz również modele tworzące obrazy i filmy (tu artefakty zobaczyć stosunkowo najłatwiej) czy analizujące stawki ubezpieczenia zdrowotnego albo niepokojące zmiany na skórze. W takich przypadkach halucynacje i uproszczenia to coś bardziej niebezpiecznego niż opowiadanie o zającach, gdy chodzi o budowę kościelnych wież.

Na tym nie koniec. Degradacja danych skutkująca „śmieciami na wyjściu" dotknie np. mniejszości (we wszystkich znaczeniach tego słowa). SI szkolona na coraz bardziej spłaszczonych, uśrednionych danych, uzna informacje dotyczące mniejszości za mniej istotne i w którymś momencie je „odetnie". To samo dotyczy mniej popularnych języków, w których tworzonych jest stosunkowo niewiele zasobów w internecie. Tu również modele SI trzeba będzie karmić danymi syntetycznymi, powielającymi błędy i redukującymi różnorodność.

Problem z całą pewnością dotknie także treści publikowanych na stronach WWW (choć tu zapewne lepiej pasuje pogardliwe określenie content) będących przeróbką, plagiatem albo nieudolną syntezą tego, co zostało napisane wcześniej. I co najgorsze, wcale nie jest powiedziane, że wcześniejsze artykuły napisał człowiek – one również mogą być dziełem sztucznej inteligencji, tyle że poprzedniej generacji. W myśl prawa Kopernika-Greshama, mówiącego, że gorszy pieniądz wypiera lepszy, treści słabej jakości, sztucznie wygenerowane, będą dominować w internecie.

„Scenariusz końca świata jest taki, że jeżeli zostawiamy to tak, jak jest przez wiele generacji, to szaleństwo zatruje dane w całym internecie" – podkreśla Richard Baraniuk, jeden z naukowców z Uniwersytetu Rice'a.

„Długoterminowe zatrucie dużych modeli językowych nie jest wcale czymś nowym. Mieliśmy już z tym do czynienia przy farmach internetowych trolli produkujących sztuczne treści, których celem było oszukanie algorytmów sieci społecznościowych oraz wyszukiwarek. Negatywne efekty, jakie przyniosły te działania, zmusiły na przykład wyszukiwarkę Google do wprowadzenia zmian – promowania treści wysokiej jakości, wytwarzanych przez źródła edukacyjne. Wyszukiwarka DuckDuckGo takie sztuczne materiały po prostu całkowicie odrzuca. To, co teraz się zmieniło wraz z powstaniem dużych modeli językowych to skala tego zatrucia" – ostrzegają naukowcy w „Nature".

Twórcy modeli SI bronią się, podkreślając, że chociaż do trenowania systemów używają danych syntetycznych, to operowanie w ten sposób przez kilka generacji w rzeczywistości się nie zdarza. „Inny problem to generalna jakość danych w internecie. Znaczna jego część to po prostu śmietnik" – powiedział w rozmowie z agencją AFP Anton Lozhkov, inżynier z Hugging Face. Jego firma przygotowując dane, odrzuca nawet 90 proc. pozyskanego z internetu materiału.

Algorytmy i ludzie

Można się również zastanawiać, na ile ten problem jest grzechem pierworodnym SI – genetycznym obciążeniem wynikającym z samej natury systemów, które określamy dziś mianem sztucznej inteligencji – i czy możemy go jakoś skorygować.

Jednym z pomysłów naprawczych jest odpowiednie przygotowanie danych dla SI w taki sposób, aby uwypuklały zjawiska rzadsze, co w teorii pozwoliłoby uniknąć ich odcięcia i pożarcia przez Uroborosa. Oznaczałoby to jednak konieczność klasyfikowania danych, ich przeróbek i rezygnację z „doskonałego" pomysłu karmienia modeli SI całym internetem.

Innym rozwiązaniem – chyba również mało realnym – jest rozróżnienie syntetycznych i rzeczywistych danych, na przykład przez specjalne znakowanie tych pierwszych. Tyle, że już obecnie internet jest wypełniony po brzegi treściami wygenerowanymi przez SI i nie są one w żaden sposób oznakowane. Zachłyśnięci możliwościami generatywnej sztucznej inteligencji przegapiliśmy moment, w którym można to było zrobić.

Najlepiej byłoby – uważa Anthon Lozhkov z Hugging Face – gdyby użytkownicy internetu po prostu nie korzystali i nie promowali sztucznie wygenerowanych materiałów. „Wierzę, że sami ludzie potrafią odróżnić efekty działania modeli SI od prawdziwych. I robią to znacznie skuteczniej niż najlepsze algorytmy" – uważa Lozhkov.

Jeszcze inny sposób zakłada odpowiednie dopasowanie wag, które są przypisywane parametrom w procesie uczenia SI w taki sposób, aby uniknąć uśredniania („odcinania końcówek"). Ale kto i w jaki sposób zmieniałby te wagi? I czy byłoby to skuteczne, biorąc pod uwagę, że w gruncie rzeczy nie wiemy dokładnie, w jaki sposób model SI dociera do ostatecznego wyniku (zjawisko czarnej skrzynki)? Pewnym wariantem tego rozwiązania jest pomysł zespołu Shumailova, aby każda kolejna generacja modelu musiała skorzystać z co najmniej 10 proc. oryginalnego zestawu danych.

Warto jednak zwrócić uwagę na zdanie, jakie pojawia się w raporcie z badań naukowców Uniwerstetu Rice'a – że bez wystarczającej ilości „świeżych, prawdziwych danych wstrzykiwanych do pętli, kolejne wersje modeli generatywnej sztucznej inteligencji skazane na produkowanie treści o coraz niższej jakości".

Te świeże i prawdziwe dane to nic innego jak oryginalna myśl ludzka. Szukajmy zatem pocieszenia w tym, że sztuczna inteligencja naprawdę nas potrzebuje – bo bez nas oszaleje.