

Demokratyzacja modeli dyfuzji: ulepszenie jakości i różnorodności za pomocą optymalizacji kierowania istniejących modeli z zachowaniem zgodności z przepisami dzięki oduczaniu

Modele dyfuzyjne stały się podstawą sztucznej inteligencji generatywnej, umożliwiając wysokiej jakości syntezę obrazów, dźwięków i tekstu. Jednak ich szerokie wdrożenie napotyka dwa kluczowe problemy: (1) istniejące modele często generują wyniki o ograniczonej różnorodności lub suboptymalnej jakości, co ogranicza ich przydatność w kreatywnych lub specjalistycznych zastosowaniach oraz (2) obawy związane z prywatnością danych, własnością intelektualną i zgodnością z przepisami (np. ustawa UE o sztucznej inteligencji) ograniczają ich wdrażanie w obszarach krytycznych lub publicznych. Te wyzwania szczególnie dotyczą użytkowników niewykwalifikowanych, małych firm i społeczności niedobrze położonych, które nie mają zasobów do trenowania modeli od zera ani do rozwiązywania problemów zgodności.

Projekt ten realizuje dwa cele, aby demokratyzować dostęp do modeli dyfuzyjnych. Po pierwsze, rozwiemy techniki kierowania, które poprawia jakość i różnorodność wyników istniejących modeli (np. Stable Diffusion) bez konieczności ich ponownego trenowania. Metody te umożliwiają lokalne dostosowanie modeli przez użytkowników niewykwalifikowanych, poprawiając rozdzielczość i kreatywność wyników. Po drugie, rozwiemy metody „zapominania” zapewniające zgodność z przepisami, dostosowane do modeli klasyfikowanych jako GPAI (UE AI Act). Skupimy się na łagodzeniu uprzedzeń i usuwaniu szkodliwych treści (np. autorskich lub niezgodnych z polityką), zachowując jakość generowanych wyników.

W zakresie optymalizacji opartej na kierowaniu zbadamy nowatorskie techniki kierowania w modelach dyfuzyjnych, które poprawia jakość i różnorodność wyników z modeli wstępnie wytrenowanych (np. Stable Diffusion). Obejmuje to modyfikacje kierowania bez klasyfikatora (classifier-free guidance) poprzez strategie adaptacyjnego skalowania oraz integrację narzędzi pomocniczych, takich jak LoRA, ControlNet lub IP-Adapter, aby kierować generowaniem treści w stronę celów zdefiniowanych przez użytkownika (np. wierność artystyczna, dokładność specyficzna dla dziedziny). Mimo że priorytetem jest minimalizowanie konieczności ponownego trenowania, skupimy się na poprawie ogólnej jakości wyników (np. rozdzielczość, spójność) i różnorodności (np. kreatywność multimodalna). Dzięki temu użytkownicy niewykwalifikowani będą mogli dostosowywać modele do zastosowań specjalistycznych – takich jak narzędzia artystyczne czy niszowe przepływy pracy – zachowując efektywność obliczeniową.

W zakresie „zapominania” zapewniającego zgodność z przepisami opracujemy strategie egzekwowania zgodności w modelach dyfuzyjnych, które są klasyfikowane jako systemy ogólnego przeznaczenia (GPAI) zgodnie z ustawą UE o sztucznej inteligencji. Prace skupia się na łagodzeniu uprzedzeń (np. stereotypowych reprezentacji) oraz zapobieganiu generowaniu treści łamiących prawa własności intelektualnej lub polityki brandingowej (np. nieautoryzowane logo korporacyjne, regulowane symbole). Poprzez celowe eliminowanie wzorców danych związanych z niezgodnymi wynikami, nasz framework „zapominania” będzie tłumić szkodliwe lub łamiące polityki generowanie, zachowując jednocześnie ogólną wydajność modelu. Praca ta bezpośrednio odpowiada na luki w modelach open-source (OSS), które przypadkowo propagują uprzedzenia lub łamia wymagania GPAI, takie jak przejrzystość i łagodzenie ryzyka w systemach o wysokim wpływie.

Łącząc innowacje techniczne z potrzebami społecznymi, projekt ten przyczyni się do:

- **Demokratyzacji dostępu:** umożliwienia użytkownikom dostosowania i personalizacji wstępnie wytrenowanych modeli lokalnie, zmniejszając zależność od intensywnego wykorzystania zasobów obliczeniowych.
- **Budowania zaufania:** zapewnienia, że systemy AI generatywnej szanują prywatność, spełniają przepisy i unikają szkodliwych uprzedzeń lub treści.
- **Zwiększenia kreatywności:** dostarczenia narzędzi artystom, edukatorom i małym przedsiębiorstwom, aby generować treści wysokiej jakości i różnorodne, dostosowane do niszowych potrzeb.

Łącząc innowacje techniczne z etycznym wdrażaniem, projekt ten zapożyczony model dyfuzyjny jako dostępne i zgodne narzędzia dla sprawiedliwego rozwoju sztucznej inteligencji.