

Rewolucyjny generatywny model syntezy głosu z firmy Meta

24 czerwca 2023

Dział badawczy amerykańskiej firmy Meta Platforms (jej działalność jest zabroniona w Rosji) ogłosił model uczenia maszynowego Voicebox zdolny do generowania języka mówionego z tekstu. Spośród innych modeli działających w tej samej dziedzinie wyróżnia się umiejętnością wykonywania wielu zadań, do których nie była szkolona, w szczególności edycji, odszumiania, narzucania stylów.



Model został wyszkolony według specjalnej metody opracowanej przez inżynierów firmy. Firma nie zamierza udostępniać Voicebox w najbliższej przyszłości z obawy przed nadużyciami.

Voicebox to generatywny model, który może syntetyzować ludzką mowę w sześciu językach: angielskim, francuskim, hiszpańskim, niemieckim, polskim i portugalskim. Jako Large Language Models (LLM) została wyszkolona do wykonywania najszerszych zadań. Jeśli LLM opiera się na statystycznych wzorcach języka i tekstu, Voicebox jest szkolony, aby nauczyć się zasad, które łączą nagrania głosowe z transkrypcją, pisze Venture Beat. Taki model można wykorzystać do rozwiązania wielu problemów przy minimalnym dostosowaniu.

„Naszym celem jest stworzenie jednego modelu, który może wykonywać wiele zadań związanych z generowaniem mowy poprzez uczenie się w kontekście” – napisali programiści w białej księdze Voicebox.

Model został wytrenowany zgodnie z metodą Flow Matching opracowaną w Meta. Jest bardziej wydajna i łatwiejsza do uogólnienia niż metody dyfuzyjne stosowane do przygotowania innych modeli. Pozwala Voiceboxowi „uczyć się z różnych danych

mowy bez starannego oznaczania tych odmian”. Bez konieczności ręcznego oznaczania danych, badacze byli w stanie trenować Voicebox z audiobookami przez łącznie zaledwie 50 tys. godzin.

Wyjątkowość Voicebox polega na tym, że może wykonywać zadania, których nie nauczono go specjalnie. Na przykład użyj dwusekundowej próbki głosu, aby wygenerować nowy tekst mówiony lub w celu wymówienia tekstu pisanego w inny sposób stylistyczny, którego próbka została jej dostarczona. A wszystko to w różnych językach. Możesz też „w locie” usuwać niepotrzebne hałasy – szczekające psy, odgłosy ruchu ulicznego i tzw. korygować rezerwacje.

Ponadto może wycinać próbki mowy z tekstu, które następnie można wykorzystać do trenowania innych modeli językowych. Testy wykazały, że ta metoda jest nie mniej skuteczna niż uczenie się z nagrań dźwiękowych mowy żywych ludzi.

Autorstwo: tallinn

Źródło: [ZmianyNaZiemi.pl](https://zmiany.naziemi.pl)