Metody optymalizacji modeli i algorytmów automatycznego rozpoznawania mowy pod kątem działania na urządzeniach mobilnych

Mikołaj Pudo

Streszczenie. Interfejsy głosowe stają się coraz bardziej popularnym sposobem komunikacji użytkownika z urządzeniami elektronicznymi. Obecnie są one już standardem w aplikacjach typu inteligentny asystent. Dane zawierające komunikaty głosowe charakteryzują się znaczną heterogenicznością. Źródła tej różnorodności mogą wystąpić już na poziomie rejestrowanego sygnału audio. Na jakość tego sygnału wpływ mają czynniki takie jak: typ mikrofonu rejestrującego dźwięk, szum tła, sposób artykulacji mówcy. Różnorodność może pojawić się również na poziomie językowym, ponieważ ta sama intencja może być wyrażona na wiele sposobów w tym samym języku. Ponadto liczba sposobów wyrażania jednej intencji znacząco się zwiększa przy rozważaniu wielu języków. Proces przetwarzania języka naturalnego przeważnie jest rozłożony na wiele etapów. W niniejszej rozprawie zaprezentowane są wyniki badań dotyczących pierwszego etapu tego procesu: konwersji mowy zawartej w strumieniu audio do zapisu tekstowego. W szczególności badania dotyczyły następujących modułów wykorzystywanych w tym procesie: wykrywanie fraz kluczowych w strumieniu audio, wykrywanie końca komendy, czy wreszcie samo dekodowanie strumienia audio do tekstu. Ponadto badania były ukierunkowane na opracowanie metod umożliwiających przeniesienie możliwie wielu wspomnianych modułów na urządzenia mobilne. Może się to odbyć poprzez opracowanie nowatorskich algorytmów lub modeli, ale również optymalizację działania istniejących rozwiązań.

W przypadku zadania wykrywania fraz kluczowych problematyczne okazało się już określenie procedury testowania tego typu rozwiązań. Dotychczasowe prace opierały się na wynikach ewaluacji przeprowadzanych na zbiorach danych, które zawierają bardzo mało fraz kluczowych, nie zawierają wymagających przykładów negatywnych lub nie są dostępne publicznie. W ramach pierwszego etapu prac przygotowano zbiór danych testowych rozwiązujący powyższe problemy i udostępniono go publicznie. Zbiór ten został nazwany Multilingual Open Custom Keyword Spotting Testset (MOCKS). Jest on oparty na publicznie dostępnych bazach audio: LibriSpeech i Mozilla Common Voice. MOCKS zawiera prawie 50 000 fraz kluczowych dla pięciu języków: angielskiego, francuskiego, hiszpańskiego, niemieckiego i włoskiego.

Kolejnym etapem badań było opracowanie metody poprawiającej skuteczność działania systemu wykrywania fraz kluczowych opartego na współczesnych modelach akustycznych. Zaproponowana metoda polega na zastosowaniu prostego modelu językowego z odpowiednio dobranymi wagami ustalanymi w fazie inicjalizacji. Przeprowadzone eksperymenty z ewaluacjami wykonanymi na zbiorze MOCKS pokazały, że możliwe jest takie ustawienie parametrów modelu językowego, które zwiększa miarę prawdziwie

pozytywną (ang. true positive rate) o około 30 punktów procentowych, z jednoczesnym wzrostem miary fałszywie negatywnej (ang. *false positive rate*) o około 20 punktów procentowych. Ponadto przeprowadzono eksperymenty z wykorzystaniem nagrań wygenerowanych przez syntezator mowy, mające na celu poprawę skuteczności modelu w przypadku specyficznych i rzadko występujących fraz kluczowych.

Efektywność systemów automatycznego rozpoznawania mowy zależy między innymi od dokładności wyznaczenia początku i końca fragmentu syganału zawierającego mowę. W szczególności istotne i skomplikowane jest wyznaczenie momentu końca frazy wypowiadanej przez użytkownika. W ramach tego problemu badawczego zaproponowano ulepszoną metodę treningu modeli wykrywających koniec komendy w strumieniu audio. Innowacją jest tu rozszerzenie standardowej funkcji kosztu o zestaw wag przypisanych do różnych części danych audio. W trakcie badań zaproponowano również metodę doboru tych wag. Przeprowadzone eksperymenty potwierdzają skuteczność opracowanej funkcji kosztu w porównaniu ze standardowym sposobem treningu modeli.

Dekodowanie strumienia audio do tekstu jest przeważnie ostatnim etapem konwersji mowy do zapisu słownego. W trakcie treningu modele stosowane w tym zadaniu wymagają dużych ilości wysokiej jakości danych. Łatwo dostępne są obszerne bazy danych audio, natomiast proces ich anotacji przeważnie jest robiony ręcznie. Z tego powodu jest on kosztowny i czasochłonny. Metody uczenia częściowo nadzorowanego stanowią próbę rozwiązania tego problemu. We wcześniejszych pracach eksperymentalnie wykazano ich skuteczność w przypadku bardzo dużych zbiorów danych audio. Natomiast badania zaprezentowane w niniejszej rozprawie pokazują, że możliwe jest zastosowanie uczenia częściowo nadzorowanego również w przypadku niewielkich baz danych. Wykorzystując zaproponowane usprawnienia do podstawowej metody, udało się obniżyć wyrazową stopę błędów (ang. *word error rate*, WER) o 12–22 punktów procentowych (w zależności od typu zbioru audio wykorzystanego do adaptacji modelu bazowego).

Słowa kluczowe: uczenie maszynowe, głębokie sieci neuronowe, przetwarzanie mowy, przetwarzanie języka naturalnego, rozpoznawanie mowy na urządzeniu, interfejs głosowy, urządzenia mobilne, projektowanie korpusów, wykrywanie fraz kluczowych, wykrywanie końca frazy w strumieniu audio, wykrywanie aktywności głosowej w strumieniu audio, rozpoznawanie mowy, uczenie częściowo nadzorowane