



Czy Nowoczesne Sieci Hopfielda mogą uczyć się reprezentacji dźwięków mowy?

Karolina Pondel-Sycz, Piotr Bilski

Politechnika Warszawska Instytut Radioelektroniki i Technik Multimedialnych
Nowowiejska 15/19, 00-665 Warszawa
karolina.pondel.dokt@pw.edu.pl

Odporne Automatyczne Rozpoznawanie Mowy (Automatic Speech Recognition, ASR) ma kluczowe znaczenie dla zastosowań w hałaśliwych warunkach akustycznych. Dotychczasowe badania nad odpornym ASR koncentrują się na rozszerzaniu danych treningowych, maskowaniu i wykorzystaniu architektur opartych o sieci neuronowe Transformer. Rozwiązania te mają wiele zalet, ale także wady, takie jak wysoka złożoność obliczeniowa i wymóg dużego zbioru danych treningowych, zwłaszcza w przypadku uczenia nowej struktury modelu. Niniejsze badanie opisuje koncepcję wykorzystania Nowoczesnych Sieci Hopfielda (ang. Modern Hopfield Networks, MHN) w celu uzyskania odpornego ASR. MHN wykorzystują odzyskiwanie pamięci asocjacyjnej, mechanizm umożliwiający przywoływanie zaszumionych lub niekompletnych wzorców. Sieć Transformer nie robi tego natywnie, nie ma mechanizmu przywracania pełnego wzorca z częściowego sygnału i wymaga zaawansowanego rozszerzenia danych do szkolenia, aby osiągnąć podobny efekt. W niniejszej pracy zaproponowano koder sygnału mowy oparty na sieci typu MHN. Zaproponowaną architekturę oparto o strukturę modelu wykorzystującą spektrogramy dźwięków mowy, zgodną z koncepcją Audio Spectrogram Transformer (AST), w której sieci Transformer zastąpiono sieciami MHN. Przedstawiono wstępne wyniki treningu oraz testów odporności modelu na szum. Zaproponowaną architekturę poddano analizie w zadaniu rozpoznawania słów z wykorzystaniem perceptronu wielowarstwowego (ang. Multilayer Perceptron, MLP). Reprezentacje sygnału mowy generowane przez sieć MHN oceniono na podstawie klasyfikacji komend głosowych z anglojęzycznej bazy danych Speech Commands V2. Wstępne wyniki szkolenia modelu na tym zbiorze danych wskazują na skuteczność na poziomie 80%.