Charakteryzacja struktury neuronowych reprezentacji ukrytych

Sieci neuronowe, obecnie coraz ważniejsza klasa algorytmów uczących się, są trenowane przez minimalizację funkcji opisującej koszt błędnych predykcji. Wartości tej funkcji zależą miedzy innymi od wag połączeń w sieci. Niestety, funkcji tej nie możemy zminimalizować analitycznie. Oznacza to, że nie mamy równania na „poprawne" wartości wag w sieci neuronowej. Funkcję kosztu minimalizujemy więc numerycznie, w procesie iteracyjnego poprawiania wag do momentu, gdy sieć osiągnie zadowalającą dokładność. Numeryczna optymalizacja funkcji kosztu powoduje, że dwie sieci wytrenowane do rozwiązywania tego samego zadania będą zazwyczaj miały różne wartości wag, nawet jeśli ich końcowa skuteczność jest bardzo zbliżona. Rodzi to następujące, istotne pytanie w pracach badawczych nad sieciami neuronowymi: czy te dwie niezależnie trenowane sieci są podobne biorąc pod uwagę sposób konstruowania predykcji? A może dwie niezależnie wytrenowane sieci (rozwiązujące dokładnie to samo zadanie) mają zupełnie odmienne wewnętrzne „zasady" konstruowania predykcji? To i podobne pytania są głównym przedmiotem badań w dziedzinie uczenia maszynowego zajmującej się podobieństwem neuronowych reprezentacji danych. Mimo że ten obszar badań ma fundamentalne znaczenie dla zastosowań sieci neuronowych, choćby w systemach, w których wymagamy objaśniania decyzji algorytmicznych, wstępne rezultaty naukowe prowadzące do odpowiedzi na podstawowe pytania dotyczące podobieństwa reprezentacje neuronowych uzyskano dopiero niedawno.

W niniejszym projekcie naukowym chcemy rozwijać metody analizy reprezentacji neuronowych wykraczające poza zagadnienie porównania dwóch wytrenowanych sieci. W tym celu planujemy zastosować dwa uzupełniające się podejścia. W pierwszej kolejności planujemy zbadać modele probabilistyczne dla reprezentacji neuronowych. W szczególności, naszym celem jest opracowanie praktycznie użytecznych modeli gęstości prawdopodobieństwa dla reprezentacji konstruowanych przez sieci neuronowe, a następnie zastosowanie tych modeli do charakteryzacji struktury w zbiorach wyuczonych reprezentacji. Po drugie, planujemy wykorzystać metody z obszaru topologicznej analizy danych do opracowania algorytmów oceny podobieństwa zbiorów reprezentacji, które koncentrują się na ich względnie niezmienniczych strukturach. Korzystając z tych dwóch podejść zaproponujemy modele i algorytmy dla szerokiej rodziny zadań w badaniach dotyczących reprezentacji neuronowych. Będziemy przykładowo pracować nad algorytmami, które wskazują dane wejściowe o unikalnej reprezentacji w sieci neuronowej, algorytmami do oceny wpływu szumu w danych uczących na konstruowane reprezentacje lub algorytmami porównującymi architektury neuronowe biorąc pod uwagę reprezentacje w wielu wytrenowanych sieciach. Nasze metody wykorzystamy również do scharakteryzowania wpływu wielu powszechnie używanych algorytmów związanych z uczeniem sieci neuronowych na reprezentacje konstruowane przez te sieci.