Głębokie uczenie dla danych tabelarycznych

Głębokie uczenie odniosło ogromny sukces w różnych dziedzinach, w tym w widzeniu komputerowym (CV), przetwarzaniu języka naturalnego (NLP) i uczeniu ze wzmocnieniem (RL). Wykorzystując dużą ilość danych i nowoczesne architektury sieci neuronowych, możemy odkryć wewnętrzne zależności w danych i nauczyć się abstrakcyjnej reprezentacji złożonych danych. W rezultacie wiele modeli głębokiego uczenia przewyższa ludzką percepcję w przypadku poszczególnych zadań, takich jak gry Atari lub rozpoznawanie obrazów.

W rzeczywistych aplikacjach najpopularniejszym typem danych są dane tabelaryczne zawierające próbki (wiersze) o tym samym zestawie cech (kolumny). Dane tabelaryczne są wykorzystywane w zastosowaniach praktycznych w wielu dziedzinach, w tym w biologii, medycynie, finansach, produkcji i wielu innych zastosowaniach opartych na relacyjnych bazach danych. Jednak modele głębokiego uczenia są przeznaczone głównie do CV i NLP, które reprezentują jedynie podzbiór rzeczywistych danych. Według ostatnich raportów programiści zajmujący się analityką danych i uczeniem maszynowym pracują z danymi tabelarycznymi równie często, jak z tekstami lub obrazami[1]. Mimo że w zastosowaniach praktycznych dominują dane tabelaryczne, modele głębokiego uczenia nie potwierdzają najwyższej wydajności w tej dziedzinie. W rzeczy samej, tradycyjne modele oparte na komitetach drzew decyzyjnych, takie jak XGBoost, pozostają podstawowym narzędziem dla większości praktyków[2].

Istnieje wiele potencjalnych powodów, dla których postęp głębokiego uczenia w CV i NLP nie znajduje odzwierciedlenia w domenie danych tabelarycznych. Nowoczesne architektury głębokiego uczenia, takie jak sieci konwolucyjne, rekurencyjne sieci neuronowe czy transformery, pojawiły się po latach badań mających na celu wykorzystanie naturalnych cech obrazów i tekstów niezmienności. **Znalezienie analogicznych niezmienników i lokalnych zależności w danych tabelarycznych jest trudne**, co sprawia, że architektury fully-connected są pierwszym wyborem w przypadku tabelarycznych zbiorów danych. Co więcej, typowe modele głębokiego uczenia zawierające miliony parametrów są trenowane na ogromnej ilości danych, co pozwala im odkrywać wyrafinowane wzorce bez przeuczenia modelu. W warunkach rzeczywistych **małe tabelaryczne zbiory danych są wszechobecne**. Jeśli wymiar danych jest stosunkowo duży w porównaniu do liczby przykładów, wówczas sieci neuronowe szybko się przeuczają, co uniemożliwia stosowanie głębszych architektur. W rezultacie dobrze ugruntowane metody drzewiaste, takie jak XGBoost lub Random Forests, są uważane za zalecaną opcję w przypadku rzeczywistych problemów z danymi tabelarycznymi.

Motywowani znaczeniem danych tabelarycznych w rzeczywistych problemach, dążymy do przeniesienia wysokiej wydajności głębokiego uczenia z CV i NLP na przypadek danych tabelarycznych.

W tym projekcie nie ograniczamy się do żadnego konkretnego zastosowania czy problemu. Skoncentrujemy się na konstruowaniu ogólnych modeli głębokiego uczenia, które można wykorzystać w różnych problemach z danymi tabelarycznymi. W szczególności skupiamy się na następujących celach:

- Budowanie głębokich modeli w celu rozwiązywania problemów o dużej liczbie przykładów.
- Projektowanie modeli głębokiego uczenia dla małych danych tabelarycznych.
- Konstruowanie słabo nadzorowanych modeli, które mogą nauczyć się reprezentacji przy minimalnej ilości informacji dostarczonych przez ludzi.
- Zwiększenie możliwości interpretacji danych tabelarycznych.

Stworzone metody i algorytmy zostaną zweryfikowane i zastosowane do różnych rzeczywistych problemów. W szczególności będziemy pracować na danych metagenomicznych, które opisują skład mikrobiomu jelitowego. Uwzględnimy także medyczne bazy danych zawierające dane pacjentów, których zadaniem jest przewidzenie obecności lub ryzyka wystąpienia określonych chorób, takich jak nowotwory, zespół jelita drażliwego i zaburzenia zdrowia psychicznego. Na koniec będziemy pracować na danych generowanych z czujników monitorujących zachowanie maszyn. W konsekwencji projekt będzie miał charakter interdyscyplinarny, a jego rezultaty będą miały wpływ także poza czystym uczeniem maszynowym.

[1] `https://www.statista.com/statistics/1241924/worldwide-software-developer-data-uses/`

[2] `https://www.kaggle.com/kaggle-survey-2021`