Streszczenie

Tytuł: *Strategie postępowania w przypadku małej liczby obserwacji uczących w problemie klasyfikacji danych nieustrukturyzowanych*

Niewystarczająca ilość danych uczących lub ich brak stanowią naturalne i często występujące wyzwania w klasyfikacji danych nieustrukturyzowanych. Nieadresowane mogą prowadzić do błędów w wynikach, co z kolei może skłonić badacza do formułowania niepoprawnych wniosków. Postęp badań przynosi jednak różne rozwiązania, mogące pomóc w rozstrzyganiu tego rodzaju problemów.

W rozprawie omówiono różne grupy metod radzenia sobie z niewielką liczbą obserwacji uczących. Zaproponowano tu również trzy autorskie podejścia: *AttentionMix* (metoda wzbogacania danych poprzez kierowane mieszanie obserwacji, dedykowana dla problemu klasyfikacji tekstów), *StatMix* (metoda wzbogacania danych, korzystająca ze statystyk zdjęcia, dedykowana dla problemu klasyfikacji obrazów w uczeniu federacyjnym) oraz standardowe uczenie sieci, oparte na ogólnodostępnych, zaszumionych danych, pobranych z internetu.

Trudności związane z klasyfikacją danych nieustrukturyzowanych są powszechnym problemem, występującym podczas budowy modeli o różnorodnym zastosowaniu (na przykład w klasyfikacji zdjęć czy analizie ilościowej komentarzy użytkowników). W takich przypadkach celem jest stworzenie modelu o jak najwyższej skuteczności w rozwiązywaniu tego typu problemów dla nowych, nieobserwowanych wcześniej danych (tj. posiadającego zdolność do generalizacji).

Rozszerzanie danych uczących stanowi jedną ze strategii regularyzacji w ramach procesu uczenia głębokich sieci neuronowych, co przekłada się na poprawę zdolności do generalizacji. Wzbogacanie danych to powszechnie stosowana strategia poprawy wyników i zwiększenia zdolności do generalizacji sieci, znajdująca szerokie zastosowanie w literaturze. W dużej mierze dominują jednak standardowe metody, takie jak obracanie w obszarze wizji komputerowej czy podmiana na synonimy w przetwarzaniu języka naturalnego. Stosunkowo niedawno pojawiły się propozycje metod, opartych na mieszaniu obserwacji.

Metody przedstawione w pracy sugerują, iż istnieje jeszcze przestrzeń do dalszych

badań w tym obszarze. Metoda *AttentionMix* pokazała, że możliwe jest dostosowanie technik mieszania do danych tekstowych, korzystając z mechanizmów, dedykowanych do przetwarzania tekstu. W ten sposób wykazano elastyczność metod mieszania w kontekście różnych modalności. Metoda *StatMix* dostosowała proces mieszania do problemu uczenia federacyjnego, gdzie istnieje potrzeba ochrony prywatności obserwacji. W tym celu wprowadzono znaczące ograniczenie ilości informacji, przesyłanych i wykorzystywanych w procesie wzbogacania danych. Wyniki przeprowadzonych eksperymentów wykazały, iż metody wzbogacania danych poprzez mieszanie obserwacji, oprócz pierwotnego zastosowania, są również elastyczne i łatwo adaptują się do innych problemów. Ich implementacja przekłada się na wzrost skuteczności uczonych modeli. Dodatkowo wykazano, że modele wysokiej jakości można skutecznie uczyć na podstawie zaszumionych danych, pobranych z internetu.

Słowa kluczowe: *wzbogacanie danych, mieszanie obserwacji, dane pobrane z internetu, wizja komputerowa, przetwarzanie języka naturalnego, uczenie federacyjne*

Dominik Lewy