Projekt - Sieci Neuronowe i Uczenie Głębokie

Projekt dotyczy przeprowadzenia badania działania Sztucznych Sieci Neuronowych dla 3 problemów – klasyfikacyjnego, regresyjnego oraz z zakresu analizy obrazów. Projekt proszę wykonać w grupach liczących od 3 do 4 osób. Projekt obejmuje przygotowanie kodów (w dowolnym języku) oraz wykonanie sprawozdania. Projekt będzie broniony (każda grupa indywidualnie) na ostatnich lub przedostatnich zajęciach. Uzyskane wyniki dodatkowo każda z grup będzie prezentować w celu ich porównania i zweryfikowania przyczyn wynikłych różnic.

Dla problemu regresyjnego i klasyfikacyjnego należy wybrać po dwie instancje problemów (1 indywidualna dla każdej grupy, natomiast druga wspólna dla całej grupy ćwiczeniowej). Na potrzeby problemu regresyjnego sugerowany jest zbiór danych zawierający szereg czasowy, gdyż pozwoli to na weryfikację możliwości wykorzystania różnych typów sieci neuronowych. Na potrzeby analizy obrazów sugerowany jest zbiór Fshion Mnist [https://www.kaggle.com/datasets/zalando-research/fashionmnist], przy czym istnieje możliwość wykorzystania innego zbioru, jeżeli wynika to np. z Państwa zainteresowania innym zagadnieniem (jak np. rozpoznawanie kurzych jaj czy wykrywaniu wody/rzek na zdjęciach).

Dane, które będziecie Państwo przetwarzać mogą dotyczyć dowolnego, interesującego dla Państwa problemu – począwszy od notowań giełdowych, walutowych, itd., przez migracje ludności i smog, a na koronawirusie czy prawdopodobieństwie wybuchu wojny skończywszy. W niniejszej części projektu, Państwa zadaniem jest prognozowanie odpowiednio wartości lub klasy dla danej obserwacji przy wykorzystaniu sieci neuronowych.

Jeżeli zamierzacie Państwo w pracy magisterskiej wykorzystywać sieci neuronowe, to istnieje możliwość połączenia projektu z obliczeniami na potrzeby pracy – kwestie z tym związane proszę ustalać indywidualnie.

Zbiory danych wykorzystane do przeprowadzenia obliczeń powinny liczyć minimum kilka tysięcy obserwacji. Warto wybierać zbiory, dla których trudniej otrzymać lepsze wyniki, gdyż wtedy łatwiej zaobserwować jest i opisać wpływ różnych parametrów sieci na otrzymywane wyniki.

Zbiory danych, które będą wspólne dla całej grupy ćwiczeniowej proszę podzielić na: zbiór uczący i testowy (w proporcji 80% do 20%) oraz na: zbiór uczący, walidacyjny i testowy (w proporcji (70%, 15%, 15%). Możecie to Państwo wykonać np. poprzez przydzielenie odpowiedniej etykiety do danej obserwacji lub fizycznym podziale danych na poszczególne pliki. Celem zdefiniowanego odgórnie podziału jest zapewnienie, aby wyniki uzyskiwane przez poszczególne grupy były do siebie porównywalne (każda grupa będzie wykonywała predykcję dokładnie dla tych samych obserwacji, stąd uzyskane wyniki będą możliwe do porównania).

Dla szeregu czasowego, proszę abyście Państwo podzielili dane na podstawie czasu (tj. pierwsze x% obserwacji stanowi próbę uczącą, następne y% próbę walidacyjną, a ostatnie z% - próbę testową). Dla danych klasyfikacyjnych, proszę aby w każdym podzbiorze danych byłby zachowane proporcje częstości występowania poszczególnych klas.

Do oceny jakości działania sieci, proszę wykorzystać kilka (3-4) miar jakości – najlepiej, aby wszystkie grupy wykorzystały te same miary, co ułatwi porównywanie wyników (np. accuracy, precision, recall, F1-score - dla klasyfikacji, oraz MSE, MAE, R^2 - dla regresji).

Sprawozdanie zawierać musi:

1. **Krótki opis** podjętych do rozwiązania problemów/zbiorów danych + linki do źródeł (lub przesłanie zbiorów danych wraz z plikami projektu).

2. **Przegląd literatury** dotyczący poruszanych problemów/zbiorów danych – konieczne jest powołanie się na inne opracowania, w których rozwiązywano dany problem, a jeżeli takich brak, zaznaczenie tego faktu i przeanalizowanie opracowań ze zbliżonej do analizowanej tematyki (pomocna strona: https://scholar.google.com). Celem tej części opracowania jest przytoczenie wyników uzyskiwanych przez inne osoby (i metod, które zostały wykorzystane) dla danego zbioru danych – co będzie stanowiło punkt odniesienia dla Państwa wyników.

Uwaga: dla zbiorów, które są wspólne dla wszystkich grup, możecie Państwo wykonać jedno opracowanie (opis + przegląd literatury), który zostanie wykorzystany przez wszystkie grupy.

3. **Analizę** wpływu wybranych parametrów na skuteczność działania sieci; w przypadku każdego parametru proszę o sprawdzenie przynajmniej 4 różnych wartości tego parametru. Do rozwiązania problemu z szeregiem czasowym proszę o wykorzystanie, poza siecią perceptronową, także sieci konwolucyjne (splotowe) i wybrany typ sieci rekurencyjnej. Dla problemu klasyfikacyjnego proszę o wykonanie obliczeń zarówno własną, jak i gotową (z wybranej biblioteki) siecią perceptronową. W przypadku analizy obrazów, poza siecią perceptronową, proszę o wykonanie obliczeń siecią konwolucyjną. Analizy parametrów powinny zawierać wnioski, które będą możliwe do weryfikacji na podstawie załączonych w sprawozdaniu tabel czy rysunków.

Lista parametrów, które należy przeanalizować: liczba warstw, liczba neuronów w warstwie, szybkość uczenia, zastosowanie optymalizatora (o ile to możliwe), wykorzystanie momementum (o ile to możliwe).

Analiza każdego z parametrów/kombinacji parametrów powinna zostać

Uwaga: Uczenie sieci nie jest procesem deterministycznym, więc dla każdego analizowanego zestawu parametrów, proces uczenia należy powtórzyć kilkakrotnie (min. 5 powtórzeń). Zawarte wyniki mogą dotyczyć wartości średnich oraz wartości najlepszych zwracanych przez sieci z zadanym zbiorem parametrów. Wykonywane zestawienia oceniające jakość sieci, powinny zawierać wyniki dla każdego z wykorzystywanych podzbiorów danych (prób: uczącej, [walidacyjnej], testowej).

4. **Podsumowanie i wnioski**

Dla zainteresowanych możliwe jest podpięcie narzędzia do optymalizacji hiperparametrów modelu (typu Optuna) i wykonanie dodatkowych analiz, co pozwoli na podniesienie oceny z projektu.

Termin wysyłania sprawozdania zostanie podany na kanale ogólnym. Za każdy rozpoczęty dzień opóźnienia maksymalna punktacja zostaje zmniejszona o 25%. Projekt wysyła jedna osoba z danej grupy (przez MS Teams). Sieci mogą być pisane w dowolnym języku/językach – sieć perceptronowa powinna być napisana bez wykorzystywania gotowych bibliotek (zawierających gotowe modele sieci). Na potrzeby własnej implementacji sieci dopuszczalne jest wykorzystywanie bibliotek numerycznych (np. numpy/pandas). Pozostałe sieci mogą być tworzone na podstawie dedykowanych bibliotek. Kody programów proszę przesłać wraz z innymi plikami projektu.

Zarówno kody, jak i samo sprawozdanie możecie Państwo wykonać w sposób umożliwiający ich późniejszą publikację (np. na GitHubie) i zamieszczenie takiej pozycji w swoim Portfolio.

Jeżeli pojawi się konieczność uściślenia wytycznych – informacje o zmianach/uściśleniach będą pojawiać się na kanale Teams.