1 Cel projektu

Szybki rozwój uczenia głębokiego spowodował masowy wzrost zastosowań głębokich sieci neuronowych w różnego rodzaju dziedzinach nauki i życia. Są one obecnie jednym z najczęstszych narzędzi m.in. w do rozpoznawania i segmentacji obrazów, tłumaczenia tekstów, odkrywaniu leków, systemach rekomendacyjnych i rozpoznawaniu mowy. Wraz z dynamicznym rozwojem uczenia głębokiego, głębokość sieci neuronowych jest coraz większa, co przekłada się na ich jakość, ale także na znacznie większy ich rozmiar oraz koszt obliczeniowy. Ten aspekt jest szczególnie problematyczny dla urządzeń mobilnych takich jak smartfony oraz dla modułów systemów wbudowanych używanych między innymi w systemach Internetu Rzeczy. Aby rozwiązać ten problem, można albo ulepszyć sprzęt komputerowy, skompresować sieci używając różnego rodzaju metod, albo poprawić wydajność obliczeniową/pamięciową modułów sieci neuronowych takich jak mechanizm atencji używany w transformerach. Ten projekt zajmuje się podejściem kompresji i poprawienia wydajności obliczeniowej/pamięciowej w głębokich sieciach neuronowych z wykorzystaniem metod aproksymacji niskiego rzędu.

2 Opis badań

Niniejszy projekt ma na celu znacząco przyczynić się do opracowania zaawansowanych metod kompresji i akceleracji splotowych i kapsułkowych sieci neuronowych oraz do zaproponowania nowych efektywnych alternatyw dla mechanizmu atencji. W projekcie stosowane będą zaawansowane metody algebry liniowej i wieloliniowej bazujące na metodach faktoryzacji macierzy i tensorów. Tensor wag warstwy splotowej może w efektywny sposób być zaproksymowany jako produkt sekwencji tensorów niższego rzędu. Z kolei moduł atencji może być zastąpiony efektywną różniczkowalną warstwą nieujemnej faktoryzacji macierzy lub tensora. Chcąc wyjść poza schematy liniowych faktoryzacji tensorów, w projekcie zostaną opracowane nowe metody faktoryzacji tensorów z przekształceniami nieliniowymi, co poprawi jakość estymowanych faktorów i pozwoli na obniżenie jeszcze niżej poziomu kompresji. Ponadto zostaną zaproponowane nowe alternatywy dla warstw atencji bazujące na projekcyjnej nieujemnej faktoryzacji macierzy oraz nieujemnej faktoryzacji tensora. W ramach eksperymentów zostaną użyte sieci różnych rozmiarów od średnich sieci na zbiorze CIFAR-10, aż do dużych sieci uczonych na zbiorze ImageNet, Pascal-VOC lub WikiText-103. Uzyskane rezultaty zostaną porównane z siecią bazową oraz najnowocześniejszymi konkurencyjnymi metodami przy użyciu metryk obliczających kompresję parametrów, kompresję operacji zmiennoprzecinkowych. Dodatkowo czas testowania oraz pobór pamięci zostanie zmierzony i porównany z czasem testowania sieci bazowej na różnych platformach obliczeniowych.

3 Spodziewane efekty

W wyniku realizacji projektu powstaną nowe metody kompresji i akceleracji głębokich sieci neuronowych bazujące na metodach aproksymacji niskiego rzędu, które z pewnością przyczynią się do dalszego rozwoju tej dziedziny. Opracowane narzędzia mogą zauważalnie usprawnić wydajność współczesnych urządzeń mobilnych oraz modułów systemów wbudowanych, poprzez skompresowanie docelowych sieci, które mają na nich działać. Naukowym spodziewanym efektem będzie przygotowanie serii czterech publikacji, które planuje się opublikować na najlepszych konferencjach związanych ze sztuczną inteligencją (np. NeurIPS, CVPR, ICCV, ECCV, ICLR, IJCAI) oraz w czasopismach takich jak IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, Knowledge Based Systems czy Neurocomputing.