Streszczenie

Kompresja splotowych sieci neuronowych przy pomocy dekompozycji tensorów

Praca doktorska koncentruje się na kompresji splotowych sieci neuronowych (CNNs). Z uwagi na to, że wagi CNNs są matematycznie reprezentowane jako czterowymiarowy tensor, możliwe jest ich kompresowanie za pomocą metod dekompozycji tensorów. W ten sposób oryginalny splot jest aproksymowany przez sekwencję mniejszych operacji. W pracy tej przedstawiono nowatorskie metody kompresji CNNs oparte na różnych dekompozycjach tensorowych. Pierwsze proponowane podejście opiera się na hierarchicznej dekompozycji Tuckera-2 (HT2), którą można postrzegać jako uogólnienie dekompozycji Tuckera-2. W podejściu HT-2 oryginalna warstwa splotowa jest zastępowana sekwencją czterech mniejszych warstw splotowych, z których dwie to sploty 1D, a pozostałe dwie to sploty jednopunktowe. Kolejne podejście opiera się na bezpośredniej dekompozycji typu tensor-train (TT), która dzięki zdekomponowaniu cyrkularnie permutowanego oryginalnego tensora wag umożliwia efektywną kompresję CNNs. W dalszej częsci pracy zaproponowano kompresję CNNs przy pomocy dekompozycji typu tensor ring (TR) o niskiej złożoności pamięciowej. Dekompozycja TR jest uogólnieniem dekompozycji TT, a jej unikalna właściwość niezmienności względem cyrkularnej permutacji umożliwia znalezienie dekompozycji TR z najniższą złożonością pamięciową przy danym dopuszczalnym błędzie, co szczególnie nadaje się do problemu kompresji sieci neuronowych. W podejściu TR, oryginalny splot jest aproksymowany przez sekwencję dwóch kontrakcji i dwóch konwolucji 1D. Ostatnie proponowane podejście opisuje ogólna technikę zagnieżdżonej kompresji CNNs, gdzie czynniki po dekompozycji są dalej dekomponowane po etapie fine-tunowania, co implikuje większą kompresję. Przeprowadzone w tej pracy eksperymenty potwierdzają efektywność proponowanych podejść, które zostały opublikowane w renomowanych i znanych czasopismach naukowych oraz materiałach konferencyjnych.

Słowa kluczowe: splotowe sieci neuronowe, kompresja sieci neuronowej, sieci tensorowe, modele dekompozycji tensorów.