УДК 004.93

Хабарлак К.С.[1]

АДАПТИВНА ПІСЛЯ НАВЧАННЯ НЕЙРОННА МЕРЕЖА

Згорткові нейронні мережі показують надзвичайно високу якість роботи у задачах комп'ютерного зору. Перші дослідження фокусувалися насамперед на якості роботи мережі, незважаючи на час її виконання. Зазвичай, такі мережі виконувалися на потужному комп'ютері завжди підключеному до живлення. Однак, велика кількість практичних застосунків [1] потребує виконання нейронних мереж на портативних пристроях, що, як правило, мають слабкий процесор та живлення від батареї, що робить дослідження архітектур мобільних нейронних мереж сучасним і затребуваним. Нейронні мережі на мобільних пристроях мають ряд переваг, такі як: відсутність потреби в безперервному доступі до інтернету – всі результати обчислюються прямо на пристрої; збереження приватності даних користувача – дані не покидають пристрій, а отже значно зменшується можливість доступу до даних третіми особами; низька затримка обробки даних та інші.

Робота із мобільними пристроями означає, що нейронну мережу необхідно виконувати на декількох поколіннях або цінових категоріях пристроїв. Як показано в роботі [2]. рівень швидкодії у них значно відмінний. В той же час сучасні архітектури нейронних мереж можуть бути сконфігуровані лише до навчання, але не після, що не дозволяє ефективно адаптувати мережу для систем із різною швидкодією. Таким чином, залишається вибір із двох варіантів: 1) навчати окрему нейронну мережу для кожної категорії пристроїв, що потребує більшого вкладання часу; 2) розробити одну архітектуру, що буде виконуватись на всіх пристроях, та в реальності буде недостатньо швидкою на бюджетних пристроях, а на потужних – буде менш точною, ніж мережа, спеціально розроблена для них. Обидва варіанти пропонують компроміс та мають значні недоліки.

В даній роботи ми пропонуємо адаптивну після навчання мобільну нейронну мережу [3], яку можна навчити один раз, а після цього конфігурувати, змінюючи швидкість її виконання. Це дозволяє динамічно адаптувати швидкість та якість мережі в залежності від потреби користувача.

Основним будівельним блоком такої мережі є Post-Train Adaptive (PTA) блок. Такий блок є простим за структурою та може бути легко

[1] Аспірант кафедри Системного аналізу та управління, Національний технічний університет «Дніпровська політехніка».

вбудованим у мобільну нейронну мережу MobileNetV2. PTA блок має декілька гілок із різною обчислювальною складністю. Гілку, яку слід виконувати може бути обрано за бажанням та під час виконання нейронної мережі, навіть прямо на мобільному пристрої. Це надає можливість конфігурувати нейронну мережу для пристроїв із значно відмінною швидкодією. Підкреслимо, що на відміну від існуючих нейронних мереж, запропонована мережа MobileNetV2 із PTA блоками навчається один раз, і вже після цього додатково конфігурується. Така конфігурація може відбуватися на основі бажаного часу, якості передбачення або ж навантаження на пристрій. Цікаво, що використання PTA блоків дозволило ще й значно покращити якість передбачення в задачі класифікації зображень, причому навіть більш швидкі у порівнянні з оригінальною MobileNetV2 конфігурації демонструють вищу точність передбачення. Зауважимо, що хоча в першу чергу така мережа націлена на виконання на мобільних пристроях, вона також може бути виконана і сконфігурована після навчання і на потужних комп'ютерах.

Висновки. На відміну від існуючих нейронних мереж, запропонована мережа MobileNetV2 із PTA блоками дозволяє ефективно змінювати архітектуру нейронної мережі після навчання, що дозволяє динамічно обирати таку конфігурацію мережі, що виконується достатньо швидко на цільовому пристрої. Цікаво й те, що запропонована мережа демонструє вищу якість за рядом метрик у порівнянні з оригінальною MobileNetV2, навіть для конфігурацій із значно меншим часом виконання. Сподіваємось, що запропонований підхід збільшить можливість використання нейронних мереж на мобільних пристроях, та дозволить краще адаптувати нейронні мережі для них.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Khabarlak K. Fast Facial Landmark Detection and Applications: A Survey / K. Khabarlak, L. Koriashkina // arXiv:2101.10808 [cs]. – 2021.

2. Хабарлак К.С. Особливості роботи методів пошуку облич на мобільних пристроях / Хабарлак К.С // Системні технології, Україна – 2021.

3. Khabarlak K. Post-Train Adaptive MobileNet for Fast Anti-Spoofing / K. Khabarlak // Proceedings of the 3nd International Workshop on Intelligent Information Technologies & Systems of Information Security with CEUR-WS, Khmelnytskyi, Ukraine, 2022: CEUR Workshop Proceedings. – CEUR-WS.org, 2022.