УДК 004.8

Костюченко А.Д. магістр спеціальності 122 Комп'ютерні науки
(Дніпровський національний університет ім. О. Гончара, м. Дніпро, Україна)

АНАЛІЗ ЕФЕКТИВНОСТІ ВИКОРИСТАННЯ АРХІТЕКТУРИ TRANSFORMER У ЗАДАЧІ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ

Класифікація зображень є однією з основних задачах комп'ютерного зору в домені штучного інтелекту, що має важливе прикладне значення в обробці медичних даних, геопросторовому аналізі, розробці критичних безпекових систем, електронній комерції. Завдання класифікації полягає у співставленні нейронною мережею об'єкта на зображенні із певним класом, відповідно до якого він належить. Зі зростанням складності зображень, що оброблюються моделлю, збільшується й обчислювальна складність та час її навчання. На початку 90-х років минулого сторіччя навчальні дані для класифікації обмежувались відносно простими чорно-білими зображеннями, при роботі з якими досить довго використовувались повнозв'язні нейронні мережі. Однак зі збільшенням складності зображень, їхніх розмірів, впровадженні кольорових каналів, кількість параметрів навчання моделі також зростає. Ефективними архітектурами, що дозволяють досягти високих показників якості, є згорткові нейронні мережі та ViT (англ. – Vision Transformer), що є відносно новим підходом до обробки зображень [1;2].

Метою дослідження була оцінка ефективності застосування моделі ViT у порівнянні зі згортковою мережею ResNet50 для задачі багатокласової класифікації з урахуванням доданих до набору даних шумів. Впровадження шуму у вхідні дані слугує корисним засобом для визначення якості навчання моделі класифікації зображень, а також надає інструменти для оцінки стабільності, стійкості до змін у даних, що використовуються у реальних задачах, та її узагальнюючої здатності.

Аналіз ефективності застосування моделі глибокого навчання Vision Transformer для розв'язку задачі класифікації проводився на навчальному наборі з веб-ресурсу Kaggle, що зазначено у [3]. Розподіл вихідного набору даних між тренувальними, валідаційними та тестовими підвибірками визначено як 70%, 15% та 15% від кількості прикладів у наборі даних. Метрика, за якою проводилась оцінка якості моделей – точність (англ. – accuracy). Шуми, що застосовувались для тестування – salt-and-pepper, гауссівський, пуассонівський, рівномірний. Результати навчання моделей, а також їх тестування на даних із шумом наведено у табл. 1-3.

Таблиця 1

Точність моделей на тренувальних, валідаційних та тестових вибірках

	Тренувальні дані	Валідаційні дані	Тестові дані
ResNet50	97,54%	98,92%	99,28%
ViT	99,3%	99,52%	99,4%

Відповідно до даних з табл. 1, обидві моделі мають високі значення точності на тренувальних, валідаційних та тестових даних.

Таблиця 2

Точність згорткової моделі на даних з шумом

ResNet50	Тренувальні дані	Валідаційні дані	Тестові дані
Salt-and-pepper шум	82,77%	81,64%	82,57%
Гауссівський шум	94%	92,96%	92,97%
Пуассонівський шум	95,27%	94,63%	93,3%
Рівномірний шум	96,93%	97,25%	96,97%

За даними з табл. 2 спостерігаємо зменшення точності моделі ResNet50 при використанні salt-and-pepper шуму на тренувальних, валідаційних та тестових даних. Незначне зменшення точності (в межах 1,5%-6%) також спостерігається і для інших видів шуму.

Таблиця 3

Точність моделі Vision Transformer на даних з шумом

ViT	Тренувальні дані	Валідаційні дані	Тестові дані
Salt-and-pepper шум	98,64%	98,68%	98,44%
Гауссівський шум	98,87%	98,68%	98,68%
Пуассонівський шум	98,7%	98,92%	98,57%
Рівномірний шум	99,26%	99,16%	99,52%

Відповідно до табл. 3, зміна точності для моделі ViT становить менше 1% для будь-якого виду шуму на тренувальних, валідаційних та тестових даних. Це свідчить про те, що використання Vision Transformer з точки зору якості моделі є ефективним підходом до вирішення задачі багатокласової класифікації, особливо при наявності додаткового шуму на зображеннях, що є важливим при використанні у реальних просторових задачах.

Список використаних джерел

1. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. [Електронний ресурс]. Режим доступу: https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

2. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale //arXiv preprint arXiv:2010.11929. – 2020. [Електронний ресурс]. Режим доступу: https://arxiv.org/pdf/2010.11929

3. Kaggle Datasets. Vehicle Image Classification [Електронний ресурс]. Режим доступу: https://www.kaggle.com/datasets/mohamedmaher5/vehicle-classification