

ЗОРОВІ ТРАНСФОРМЕРИ В ЕНДОСКОПІЇ

Вінницький національний технічний університет

Анотація: В даній роботі проведено огляд архітектури зорового трансформера (ЗТ), що є новітньою альтернативою традиційним згортковим нейронним мережам (ЗНМ) у галузі комп'ютерного зору. Проаналізовано фундаментальні відмінності між ЗТ та ЗНМ, зокрема відсутність у трансформерів вбудованих індуктивних упереджень, таких як локальність та трансляційна еквівалентність, що компенсується масштабним попереднім навчанням на великих наборах даних. Досліджено ефективність застосування ЗТ для аналізу медичних зображень, де вони демонструють переваги у задачах з обмеженою кількістю даних та складними, неочевидними патернами. Окремо висвітлено тенденцію до спрощення архітектур та переходу до моделей типу "encoder-only" (ЕоМТ), які, відмовляючись від складних спеціалізованих компонентів (адаптерів, декодерів), досягають високої точності та значно вищої швидкості обробки, що є критично важливим для клінічного застосування в ендоскопії.

Ключові слова: *ШІ, машинне навчання, нейронні мережі, зорові трансформери, згорткові нейронні мережі, ендоскопія, ResNet, капсульна ендоскопія, залишкова нейронна мережа*

Вступ

Згорткові нейронні мережі протягом тривалого часу були домінуючою архітектурою в комп'ютерному зорі, особливо в аналізі медичних зображень [1]. Такі моделі, як ResNet та EfficientNet, довели свою високу ефективність у задачах класифікації та сегментації завдяки вбудованим індуктивним упередженням - локальності обробки та трансляційній еквівалентності, які природно відповідають структурі зображень [2]. Проте успіхи архітектури трансформера в обробці природної мови (NLP) стимулювали її адаптацію для застосування у обробці і аналізі зображень. Результатом цього став зоровий трансформер - модель, що кидає виклик парадигмі згортковим нейронним мережам. ЗТ обробляє зображення, розбиваючи його на послідовність невеликих ділянок (патчів) і аналізуючи їх як послідовність токенів за допомогою механізму самоуваги. Цей підхід дозволяє моделювати глобальні залежності між будь-якими частинами зображення вже на ранніх шарах мережі, що є принциповою відмінністю від локальних операцій згортки в ЗНМ.

Результати дослідження

Критично важливим для ефективного застосування моделей глибокого навчання в медичній діагностиці є глибоке розуміння їхніх фундаментальних архітектурних відмінностей. Вибір між ЗНМ та ЗТ може суттєво вплинути на точність, надійність та обчислювальну ефективність діагностичних систем, що в свою чергу обумовлене підробом мережі під поставлені задачі перед розробниками.

ЗНМ функціонують на основі згорткових шарів, які застосовують фільтри до локальних ділянок зображення. У саму архітектуру ЗНМ вбудовані так звані індуктивні упередження (inductive biases), такі як локальність (припущення, що сусідні пікселі пов'язані між собою) та інваріантність до зсуву (здатність розпізнавати об'єкт незалежно від його положення). Це дозволяє ЗНМ ієрархічно виділяти ознаки — від простих (краї, текстури) на початкових шарах до складних (форми, об'єкти) на глибших [1]. На противагу цьому, ЗТ запозичують підхід з обробки природної мови. Вони розбивають зображення на сітку невеликих ділянок-патчів і обробляють їх як послідовність, аналогічну до слів у реченні [1]. Завдяки механізму самоуваги, що лежить в основі стандартного кодувальника трансформера [3], ЗТ здатні моделювати глобальні взаємозв'язки між усіма патчами одночасно, починаючи з найпершого шару мережі.

Це дозволяє їм "бачити" зображення цілісно, а не по частинах, що є фундаментальною відмінністю від локального підходу CNN[1]. Архітектура зорового трансформера представлена на рисунку 1.

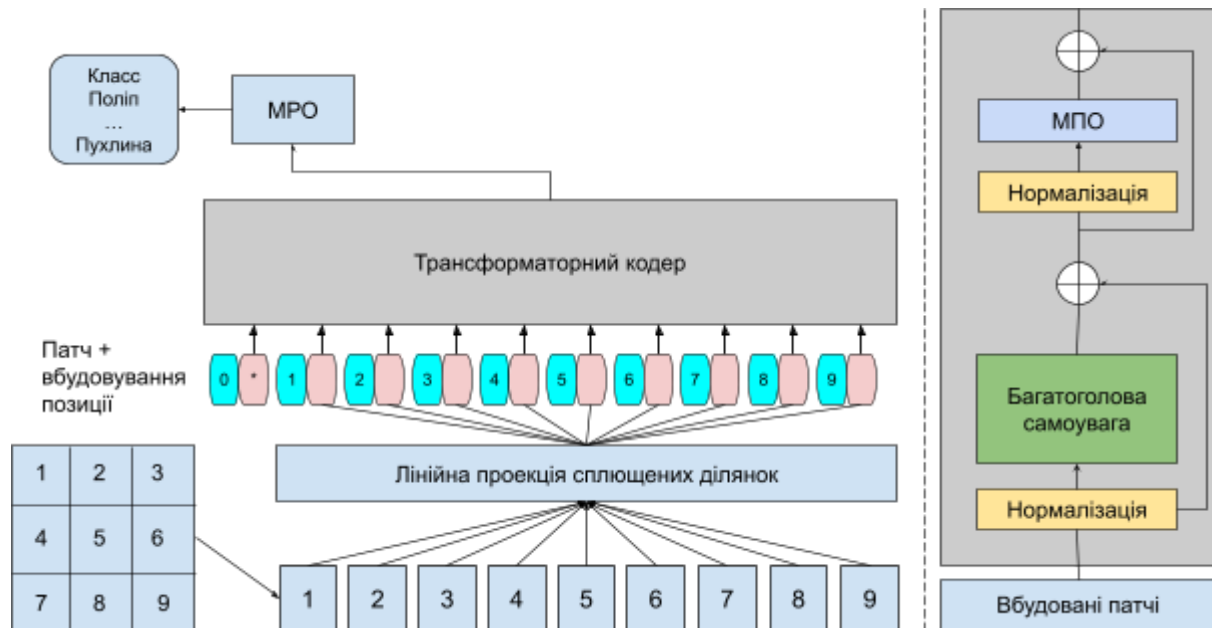


Рисунок 1. Архітектура зорового трансформера

Порівняльний аналіз показує, що вибір оптимальної архітектури значною мірою залежить від специфіки завдання, розміру набору даних та характеристик зображень [4].

Діагностика пневмонії (рентгенівські знімки): Архітектури ЗНМ, зокрема ResNet-50, продемонстрували найвищу точність (98.37%). Це підтверджує їхню стабільну ефективність для великомасштабного аналізу рентгенограм, де вбудовані індуктивні упередження допомагають ефективно виявляти патологічні зміни [4].

Класифікація пухлин мозку (МРТ): Хоча Vision Transformers зазвичай потребують великомасштабного попереднього навчання для досягнення високої продуктивності на невеликих наборах даних [1], модель ViT DeiT-Small продемонструвала виняткові результати (92.16%) на найменшому датасеті в цьому аналізі. Це свідчить про те, що глобальний механізм уваги ViT [1] особливо ефективний для виявлення ледь помітних, просторово розподілених патологічних ознак, характерних для МРТ-зображень пухлин мозку. В цьому сценарії перевага глобального контексту виявляється важливішою за слабші індуктивні упередження [4].

Аналіз на наявність меланоми (дерматоскопія): Найкращий результат (81.84%) показали ефективні CNN, такі як EfficientNet-B0. Це може вказувати на перевагу CNN у задачах, що вимагають аналізу дрібнозернистих текстур та локальних патернів, характерних для дерматоскопічних зображень [4].

Висновки

Дана робота підкреслює, що, зорові трансформери є потужною технологією для аналізу ендоскопічних зображень. Його здатність моделювати глобальні залежності та ефективність на малих наборах даних після масштабного попереднього навчання відкривають нові можливості для виявлення складних патологій, що також впливає і на покращення і дослідження цього підходу у використанні в ендоскопії. Тенденція до спрощених "encoder-only" архітектур обіцяє створення високоточних та швидких діагностичних інструментів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Версія 2). arXiv. <https://doi.org/10.48550/ARXIV.2010.11929>
2. Пруданєн, Ю., & Кожем'яко, А. (2025). Класифікація зображень із застосуванням оптико-цифрових методів покращення якості зображень та глибокого навчання при проведенні ендоскопічних досліджень. Оптико-електронні інформаційно-енергетичні технології, 49(1), 135–146. <https://doi.org/10.31649/1681-7893-2025-49-1-135-146>
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need (Версія 7). arXiv. <https://doi.org/10.48550/ARXIV.1706.03762>
4. Kawadkar, K. (2025). Comparative Analysis of Vision Transformers and Convolutional Neural Networks for Medical Image Classification (Версія 1). arXiv. <https://doi.org/10.48550/ARXIV.2507.21156>

Пруданєн Юрій Євгенович — аспірант кафедри БМІОЕС, факультет інформаційних електронних систем, Вінницький національний технічний університет, Вінниця, e-mail: mcpchip@gmail.com;

Кожем'яко Андрій Вікторович — кандидат техн. наук, доцент кафедри біомедичної інженерії та оптико-електронних систем, Вінницький національний технічний університет, м. Вінниця.

VISUAL TRANSFORMERS IN ENDOSCOPY

Vinnitsia National Technical University

Abstract: This paper reviews the architecture of visual transformers (VT), which are a novel alternative to traditional convolutional neural networks (CNN) in the field of computer vision. The fundamental differences between VT and CNN are analyzed, in particular, the absence of built-in inductive biases in transformers, such as locality and translation invariance, which are compensated for by large-scale pre-training on large datasets. The effectiveness of using transformers for medical image analysis is investigated, where they demonstrate advantages in tasks with limited data and complex, non-obvious patterns. The trend towards simplifying architectures and transitioning to encoder-only models (EoMT) is highlighted separately. By abandoning complex specialized components (adapters, decoders), these models achieve high accuracy and significantly higher processing speeds, which is critical for clinical applications in endoscopy.

Keywords: AI, machine learning, neural networks, visual transformers, convolutional neural networks, endoscopy, ResNet, capsule endoscopy, residual neural network

Yurii Yevhenovych Poudanien – Post-Graduate student of the Biomedical Engineering, e-mail: mcpchip@gmail.com.

Kozhemiako Andrii Viktorovich — Candidate of Engineering Sciences, docent of Biomedical Engineering and Optical-Electronic Systems, Vinnitsia National Technical University, Vinnitsia.