УДК 621.38:004.93.032.2

МЕТОДИ ПІДВИЩЕННЯ ТОЧНОСТІ ТА ЕФЕКТИВНОСТІ СИСТЕМ РОЗПІЗНАВАННЯ ОБРАЗІВ НА ОСНОВІ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

Кавецький В.С., Стрілкова Т.О.
e-mail: viacheslav.kavetskyi@nure.ua, tetiana.strilkova@nure.ua
Харківський національний університет радіоелектроніки, каф. МЕЕПП
м. Харків, Україна

This work overviews the evolution of modern computer vision from hand-crafted feature descriptors to deep hierarchical models that automatically learn complex visual representations. Convolutional neural networks construct multi-level feature hierarchies, complicating small object detection. Feature pyramid networks address scale variance by combining semantic richness with multi-resolution maps. Fully convolutional networks with skip connections fuse deep semantics and details in segmentation, evaluated by mean Intersection over Union.

Сучасний етап розвитку систем розпізнавання образів характеризується відмовою від використання ручних дескрипторів ознак. Активніше використовуються глибокі ієрархічні моделі, здатні автоматично формувати складні ознаки. Згорткові нейронні мережі створюють багаторівневі ієрархії ознак, у яких початкові шари реагують на низькорівневі структури, тоді як глибші шари захоплюють семантично значущі структури [1]. Водночас стандартні архітектури стикаються з суттєвою проблемою просторової варіативності – операції субдискретизації призводять до швидкої втрати просторової інформації, через що дрібні об’єкти не виявляються повністю [2].

Метою роботи є розробка підходу та методу, який забезпечує високу точність і надійність розпізнавання образів, здатний адаптуватися до складних умов та масштабуватися для обробки великих обсягів даних.

Одним із найефективніших способів подолання цієї проблеми стало впровадження мережі пірамід ознак. У таких системах будуються набори ознак з високою семантичну вагою та достатньою роздільною здатністю. Процес призначення регіону інтересу до відповідного рівня ієрархії ознак визначається через математичну залежність:

$$k = \left\lfloor k_0 + \left(\sqrt{\frac{wh}{224}} \right) \right\rfloor$$

де w та h – ширина регіонів,
k_0 – базовий рівень.

Використання константи 224 гарантує, що об’єкт з площею 224^2 буде оброблятися на рівні k_0, а об’єкти іншого розміру будуть перенаправлятися на відповідні рівні з іншою роздільною здатністю.

У завданнях семантичної сегментації та локалізації виникає суперечливість між потребою в глибокій семантиці та збереженням високої просторової точності. Повністю згорткові мережі вирішують це через впровадження архітектури з пропускними зв'язками з метою поєднання семантичної інформації з детальною інформацією про вигляд об'єкту. Ефективність таких систем оцінюється через метрику середнього перетину над об'єднанням, яка обчислюється як:

$$IU = \left(\frac{1}{n_{cl}}\right)\sum_{i} n_{ii} \,/\, (t_i + \sum_{j} n_{ji} - n_{ii}),$$

де n_{ij} – кількість пікселів клас i ідентифікованих як клас j,
n_{cl} – кількість пікселів, які належать до інших класів,
t_i – загальна кількість пікселів класу i.

У системах визначення об'єктів в реальному часі задача розпізнавання переходить у єдину задачу регресії, де імовірність класу та координат стробу обчислюється одночасно для всього зображення. У моделях типових до Fast R-CNN для підвищення точності застосовується функція втрат:

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1]L_{loc}(t^u, v),$$

де L_{cls} – відповідає за помилкове визначення класу об'єкта,
L_{loc} – мінімізує похибку координат стробу.
λ – регулює баланс між цими завданнями, дозволяючи системі навчатися спільно.

Сучасні системи аналізу сцен переходять до моделювання складних ієрархічних зв'язків для оцінки взаємодії між суб'єктом, виконуваною дією та цільовим об'єктом. Замість поетапного виділення кожного елемента, архітектура обчислює ймовірність їх спільного існування в межах конкретної сцени. Використовується концепція просторової сумісності, яка базується на припущені про вигляд та взаємодії людини, – це надзвичайно інформативний сигнал для локалізації об'єктів.

Такий підхід дозволяє ієрархічним мережам ефективно ідентифікувати об'єкти навіть у складних умовах, наприклад, при їх частковому перекритті іншими предметами або за низької якості вхідного сигналу. Візуальні ознаки людини, що перебуває в позі сидіння, дозволяють моделі заздалегідь очікувати наявність стільця або дивана під суб'єктом, що звужує простір пошуку об'єктів та мінімізує помилкові спрацювання. Таким чином, процес цифрової обробки зображень перетворюється із пасивного аналізу пікселів до активного прогнозування положення об'єктів [3].

Підвищення складності моделі та її здатність до генерації розгорнутих описів сцен породжує проблему галюцинацій. Для боротьби з цим дефектом та для перевірки надійності розпізнавання об'єктів впроваджується спеціалізована метрика оцінки релевантності [4]. Вона кількісно визначає частку неіснуючих об'єктів у вихідних даних моделі відносно реального

візуального вмісту, зафіксованого еталонною розміткою. Використання такої метрики необхідно для підтвердження, що робота системи базується на справжніх візуальних ознаках.

Оптимізація процесу навчання складних систем в умовах малої кількості розмічених даних досягається впровадженням методів активного навчання [3]. Використання змагальних стратегій у латентному просторі дозволяє моделі самостійно визначати найбільш інформативні зразки для подальшої розмітки [4]. Такий підхід дозволяє підтримувати високу точність розпізнавання об'єктів при невеликій ручній розмітці, роблячи цифрові системи здатними до масштабування в умовах обробки великих відеопотоків.

Сучасні системи розпізнавання образів досягли високого рівня точності. Подальший розвиток галузі пов'язаний із впровадженням мереж, здатних адаптувати свою структуру під логіку конкретного запиту та додаванням часового маркування.

Список використаних джерел:

1. Object Hallucination in Image Captioning / A. Rohrbach et al. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Stroudsburg, PA, USA, 2018.

2. Localizing Moments in Video with Natural Language / L. A. Hendricks et al. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 22–29 October 2017. 2017.

3. Sinha S., Ebrahimi S., Darrell T. Variational Adversarial Active Learning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October – 2 November 2019. 2019.

4. Deformable part models are convolutional neural networks / R. Girshick et al. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. 2015.

5. Стрілкова Т. О. Литюга О.П., Дуднік О.В. Використання методів машинного навчання для інтелектуального аналізу даних систем відеоспостереження // XXIII Міжнародна науково-технічна конференція "ПРИЛАДОБУДУВАННЯ: стан і перспективи". Секція 2. Оптичні та оптико-електронні прилади системи. Фотоніка. 14-15 травня 2024 року, КПІ ім. Ігоря Сікорського, Київ, Україна., С. 51-52.