ДОСЛІДЖЕННЯ ПІДХОДІВ ГЛИБИННОГО НАВЧАННЯ ДЛЯ РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ ТА ОБРОБКИ ПРИРОДНОЇ МОВИ: КОНЦЕПТ ДЛЯ СИСТЕМИ Є-ВОРОГ

Владислав Терницький, Дмитро Широкорад
Національний університет «Запорізька політехніка»

EXPLORING DEEP LEARNING APPROACHES FOR IMAGE RECOGNITION AND NATURAL LANGUAGE PROCESSING: THE CONCEPT FOR E-ENEMY

Abstract. The large-scale invasion of Ukraine by Russia has highlighted the need for innovative approaches to gather and process information in war zones. With the rapid growth of artificial intelligence (AI) models and natural language processing (NLP) techniques, the concept of leveraging deep learning approaches for improving the accuracy and efficiency of intelligence gathering becomes increasingly relevant. This paper explores the concept for e-Enemy, a system developed by the Ukrainian government to gather information on enemy movements through user-submitted applications. The main goal of the concept is to evaluate the accuracy of images and text submitted in the applications to the e-Enemy, the accuracy of text-image correspondence and ensure filtering spam applications. This research aims to contribute to the development of more accurate and efficient methods for extracting valuable intelligence from user-generated content within the context of the e-Enemy system.

Розпізнавання об'єктів відіграє ключову роль у концепті для системи єВорог, оскільки дозволяє виявляти та ідентифікувати ворожу техніку або живу силу на зображеннях, надісланих користувачами. Для вирішення цього завдання було обрано модель You Only Look Once (YOLO) [1] через її ефективність у виявленні та локалізації об'єктів. Вибір даної моделі мотивований її здатністю досягати балансу між точністю та швидкістю обробки. Як наскрізна архітектура згорткової нейронної мережі (CNN) [2], YOLO обробляє все зображення за один прохід і одночасно прогнозує обмежувальні рамки та ймовірності класів для декількох об'єктів. Модель використовує уніфіковану архітектуру, яка розділяє вхідне зображення на сітку і генерує прогнози на основі ознак, виділених у кожній комірці сітки. Вона складається з основної мережі CNN, як правило, варіанту Darknet, за якою слідують кілька згорткових шарів, що прогнозують координати обмежувальних рамок і пов'язані з ними ймовірності класів. Система моделює виявлення як регресійну задачу. Вона розбиває зображення на сітку $S \times S$ і для кожної

комірки прогнозує B клітинок, довірчу ймовірність для цих клітинок, та ймовірності класів C. Ці прогнози кодуються у вигляді $S \times S \times (B * 5 + C)$ тензор [1].

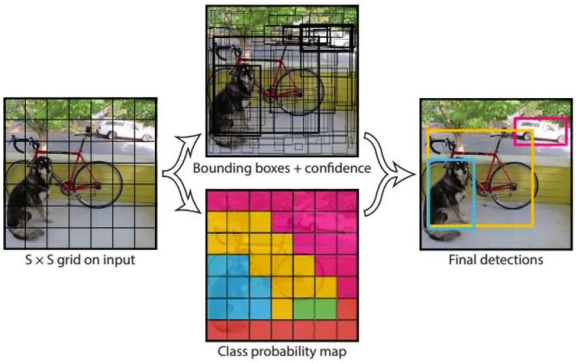

Рис. 1 - Модель YOLO [1]

YOLO використовує єдину мережеву оцінку для прогнозування граничної області, що робить її значно швидшою порівняно з іншими моделями виявлення об'єктів. Крім того, YOLO відома своєю здатністю обробляти невеликі об'єкти завдяки притаманній їй просторовій чутливості та контекстному розумінню.

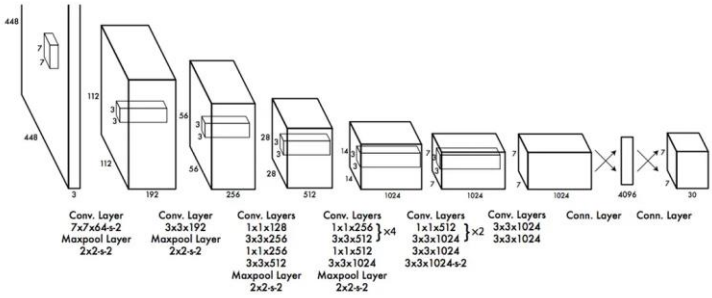

Рис.1 - Архітектура моделі YOLO [1]

Мережа виявлення має 24 згорнуті шари, за якими слідують 2 повністю з'єднані шари. Згорткові шари 1 × 1, що чергуються, зменшують простір ознак від попередніх шарів.

Основні переваги YOLO:

- модель досягає вражаючої продуктивності в режимі реального часу, що дозволяє оперативно аналізувати зображення, надіслані користувачем. Це має вирішальне значення в контексті системи єВорог, де своєчасне і точне виявлення ворожої техніки або живої сили є життєво важливим для прийняття стратегічних рішень і безпеки постраждалого населення;
- уніфікована архітектура YOLO дозволяє ефективно фіксувати контекстні зв'язки між об'єктами на зображенні, що полегшує точну ідентифікацію і локалізацію декількох об'єктів одночасно. Ця можливість є особливо корисною у сценаріях, коли необхідно виявити декілька одиниць ворожої техніки або живої сили на зображенні.

Проте важливо зазначити, що модель YOLO може зіткнутися з проблемами при роботі з сильно закритими або перекритими об'єктами, оскільки вона покладається на комірки сітки для генерування прогнозів. Крім того, через однопрохідну природу YOLO, вона може мати обмеження в точному виявленні об'єктів різного масштабу на зображенні.

У контексті обробки тексту, поданого разом із зображеннями, та перевірки відповідності між текстом і зображенням, було вирішено застосувати модель BERT (Bidirectional Encoder Representations from Transformers) [3] для задач обробки природної мови (NLP). Модель BERT використовує transformer архітектуру, яка складається з декількох шарів з механізмами самоуваги. Ця архітектура дозволяє моделі фіксувати залежності та зв'язки між словами в реченні, тим самим забезпечуючи глибше розуміння тексту. Двонаправлений характер BERT та глибоке розуміння контексту роблять його ефективним для розпізнавання тексту в системі єВорог. Використовуючи BERT, ми зможемо вловити складні значення та нюанси, присутні в текстових описах, пов'язаних з надісланими зображеннями.

У нашому конкретному випадку використання модель BERT буде відповідати за розпізнавання та розуміння текстової інформації, наданої користувачами у поєднанні з надісланими зображеннями. Використовуючи попередньо навчену модель, основною ціллю буде отримувати з тексту змістовне представлення інформації, яке можна використати для оцінки відповідності між текстовими описами та візуальним контентом, зображеним на відповідних фотографіях.

Однією з помітних переваг моделі BERT є її здатність генерувати та розпізнавати зв'язну та контекстуально багату текстову інформацію. Ця здатність необхідна для оцінки того, наскільки точно текст відображає візуальний контент - ворожу техніку або живу силу. Використовуючи здатність BERT фіксувати семантичні зв'язки і контекстну інформацію,

ми можемо проаналізувати відповідність між текстом і зображенням і визначити рівень відповідності або невідповідності між цими двома модальностями. Однак доцільно звернути увагу, що для забезпечення високої точності розпізнавання і розуміння тексту може знадобитися додаткове донавчання моделі використовуючи розмічений набір даних з текстовими описами.

Загальний алгоритм оцінки відповідності між текстом та візуальним контентом буде виглядати наступним чином:

1. Виділення ознак (features). Отримання ознак зображень з вихідних обмежувальних рамок та текстових ознак з текстових описів.

2. Порівняння та відповідність. Обчислення показника подібності або метрики відстані між елементами зображення і тексту, наприклад, косинусової подібності [4] або евклідової відстані [5]. Цей показник вказує на ступінь відповідності між текстовим описом і виявленим візуальним контентом, що дає змогу кількісно оцінити схожість або відмінність між елементами.

3. Застосування порогу до оцінки подібності, щоб визначити, чи достатньою мірою текст і зображення відповідають один одному. Щоб визначити цей поріг, ми можемо використовувати валідаційний набір даних з анотованими парами текст-зображення. Порівнюючи оцінки подібності, отримані за допомогою алгоритму, з базовими анотаціями, ми можемо вибрати відповідне порогове значення, яке відповідає бажаному рівню точності їх відповідності.

Реалізувавши цей алгоритм, ми можемо ефективно оцінити точність відповідності текст-зображення в системі єВорог, збільшуючи загальну продуктивність роботи системи та забезпечуючи дієву фільтрацію спам заявок, а також полегшити роботу аналітиків системи.

Література

1. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, University of Washington, Allen Institute for AI, Facebook AI Research, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, http://pjreddie.com/yolo/

2. Keiron O'Shea, Ryan Nash (2015). An Introduction to Convolutional Neural Networks. https://arxiv.org/abs/1511.08458

3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. https://arxiv.org/abs/1810.04805

4. Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi, Kumamoto University. Semantic Cosine Similarity, 2012.

https://www.researchgate.net/publication/262525676_Semantic_Cosine_Similarity

5. Per-Erik Danielsson, Euclidean distance mapping, Computer Graphics and Image Processing, Volume 14, Issue 3, 1980, Pages 227-248, ISSN 0146-664X, https://doi.org/10.1016/0146-664X(80)90054-4.