МАГІСТЕРСЬКА РОБОТА

МР. ІПм - 44.00.00.000 ПЗ

Група ІПм-24-3

Максимів В`ячеслав

2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Максимів В`ячеслав Любомирович

(прізвище, ім'я, по батькові)

УДК 004.9

(індекс)

МАГІСТЕРСЬКА РОБОТА

Інтелектуальний аналіз кулінарного контенту

(назва роботи)

Інженерія програмного забезпечення

(назва освітньої програми)

121 - Інженерія програмного забезпечення

(шифр і назва спеціальності)

Максимів В.Л.

(підпис, ініціали та прізвище здобувача освітнього ступеня)

Науковий керівник Храбатин Роман Ігорович, к.ф.-м.н., доцент

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Допущено до захисту

Завідувач кафедри

доц. Бандура В.В.

(посада) (підпис) (дата) (ініціали та прізвище)

Нормоконтроль

доц. Вовк Р.Б.

(посада) (підпис) (дата) (ініціали та прізвище)

Робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Івано-Франківськ – 2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Освітній рівень магістр

Спеціальність 121 – Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:
Зав. кафедрою **ІПЗ**

доц. В.В. Бандура

“ 04 ” вересня 2025 р.

ЗАВДАННЯ

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Максиміву В`ячеславу Любомировичу

(прізвище, ім’я, по-батькові)

1. Тема магістерської роботи “ Інтелектуальний аналіз кулінарного контенту ”

керівник проекту (роботи) Храбатин Р.І., к.ф.-м.н., доцент

затверджені наказом закладу вищої освіти від “ 05 ” листопада 2025 р. № 695/7

2. Строк подання студентом проекту (роботи) 15 грудня 2025 р.

3. Вихідні дані до проекту (роботи) Концепції та формальні моделі, методи побудови інформаційних технологій інтелектуального аналізу даних

4. Зміст розрахунково - пояснювальної записки(перелік питань, які потрібно розробити)

1. Аналіз предметної області інтелектуального видобування знань з кулінарного контенту
2. Огляд досліджень в області візуального сприйняття та представлення знань у кулінарній сфер
3. Методології та рішення інтелектуального аналізу кулінарного контенту
4. Реалізація методів та методології інтелектуального аналізу кулінарного контенту

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)

1. Архітектура NELL - Never-Ending Language Learner (рис. 1.1)
2. Схема ітеративного підходу побудови структурованої БЗ для розпізнавання сцени (рис. 1.2)
3. Ілюстрація простої моделі класифікації зображень (рис. 1.3)
4. Архітектура VGG моделі мережі (рис. 1.4)
5. Архітектура Faster R-CNN (рис. 1.7)

6. Консультанти розділів проекту (роботи)

Розділ	Консультант	Підпис, дата
Перевірка на плагіат	доц., к.т.н. Вовк Р.Б.	

7. Дата видачі завдання 04 вересня 2025 р.

Керівник _____
(підпис)

Завдання прийняв до виконання_____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назви етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Підбір і вивчення літератури по темі магістерської роботи	15.09.2025	виконано
2	Аналіз предметної області інтелектуального видобування знань з кулінарного контенту	01.10.2025	виконано
3	Огляд досліджень в області візуального сприйняття та представлення знань у кулінарній сфері	17.10.2025	виконано
4	Методології та рішення інтелектуального аналізу кулінарного контенту	02.11.2025	виконано
5	Реалізація методів та методології інтелектуального аналізу контенту	19.11.2025	виконано
6	Методологія передбачення інгредієнтів багатоскладових страв на основі їх зображень	02.12.2025	виконано
7	Затвердження пояснювальної записки роботи завідувачем кафедри	15.12.2025	виконано

Студент – магістр _____
(підпис)

Керівник роботи _____
(підпис)

АНОТАЦІЯ

Магістерська робота: 85 с., 32 рис., 12 табл., 42 джерела.

Тема: Інтелектуальний аналіз кулінарного контенту

Мета магістерської роботи: розробка та обґрунтування методології інтелектуального аналізу кулінарного контенту, що поєднує глибинні моделі, графові структури знань та спеціалізовані алгоритми для розпізнавання об’єктів, станів і функціональних дій у кулінарному контенті.

Об’єктом дослідження є процеси інтелектуального аналізу кулінарного контенту.

Предметом дослідження є методи, моделі та інструменти інтелектуального аналізу і представлення знань, що забезпечують розпізнавання об’єктів, кулінарних станів, дій та процедурних залежностей у кулінарному контенті.

Результати дослідження

На основі аналізу сучасних методологій було запропоновано та обґрунтовано застосування функціональної об’єктно-орієнтованої мережі як апріорного представлення знань про кулінарні дії та об’єкти.

Висновок

В дослідженні була запропонована методика прогнозування інгредієнтів багатоскладових страв за їхнім зображенням, що свідчить про практичний потенціал систем інтелектуального аналізу у сфері харчових технологій.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ, КУЛІНАРНИЙ КОНТЕНТ, ГЛИБИННЕ НАВЧАННЯ, ГРАФОВІ МОДЕЛІ ЗНАНЬ, ФУНКЦІОНАЛЬНІ ОБ’ЄКТИ, ФУНКЦІОНАЛЬНІ ДІЇ, РОЗПІЗНАВАННЯ СТАНІВ, КЛАСИФІКАЦІЯ СТРАВ.

ABSTRACT

Master Thesis: 85 pp., 32 fig., 12 tab., 42 sources.

Topic: Intelligent analysis of culinary content

The purpose of the master's thesis: development and justification of the methodology of intelligent analysis of culinary content, which combines deep models, graph knowledge structures and specialized algorithms for recognizing objects, states and functional actions in culinary content.

The object of the study is the processes of intelligent analysis of culinary content.

The subject of the study is methods, models and tools of intelligent analysis and knowledge representation that ensure the recognition of objects, culinary states, actions and procedural dependencies in culinary content.

Research results

Based on the analysis of modern methodologies, the use of a functional object-oriented network as an a priori representation of knowledge about culinary actions and objects was proposed and substantiated.

Conclusion

The study proposed a methodology for predicting the ingredients of multi-ingredient dishes based on their image, which indicates the practical potential of intelligent analysis systems in the field of food technology.

INTELLIGENT ANALYSIS, CULINARY CONTENT, DEEP LEARNING, GRAPHIC KNOWLEDGE MODELS, FUNCTIONAL OBJECTS, FUNCTIONAL ACTIONS, STATE RECOGNITION, DISH CLASSIFICATION

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

FOON - Functional Object Oriented Network

ResNet - Residual Network

R-CNN - Region-based Convolutional Neural Network

DCN - Deep Convolutional Network

RNN - Recurrent Neural Network

C-SID - Cooking State Identification Dataset

MTurk - Amazon Mechanical Turk

ReLU - Rectified Linear Unit

BCE - Binary Cross-Entropy

ВСТУП

Актуальність теми.

Сучасний розвиток штучного інтелекту, зокрема комп'ютерного зору, глибинного навчання та графових моделей знань, створює передумови для автоматизації складних когнітивних процесів у різноманітних прикладних сферах. Одним із таких напрямів є інтелектуальний аналіз кулінарного контенту, що охоплює опрацювання зображень, відео, рецептів та структурованих знань про кулінарні дії, інгредієнти й технологічні процеси. Кулінарний контент, представлений у вигляді інструкцій, відеоуроків та фото готових страв, містить значний обсяг латентної інформації, здатної бути формалізованою та використаною у рекомендаційних системах, робототехнічних комплексах та інтелектуальних асистентах. Водночас структура кулінарних даних є багатомодальною: для їхнього якісного аналізу необхідно одночасно обробляти візуальні властивості інгредієнтів, функціональні дії, часові послідовності та причинно-наслідкові зв'язки між об'єктами.

Попри значний прогрес у галузі комп'ютерного зору, існує потреба у методах, здатних забезпечити глибоке семантичне розуміння кулінарних процесів, а не лише базове розпізнавання об'єктів. Класичні нейромережеві архітектури демонструють високу точність у задачах класифікації зображень, але їхня здатність до інтерпретації складних рецептурних процедур є обмеженою. Саме тому актуальним є інтеграція графових моделей знань, які дозволяють структурувати інформацію про кулінарні об'єкти, стани, рухи та функціональні відношення між ними. Це створює підґрунтя для побудови систем, здатних аналізувати кулінарні відео у вигляді послідовності логічно взаємопов'язаних дій і формувати високорівневе представлення рецепту.

У магістерській роботі здійснюється комплексний підхід до інтелектуального аналізу кулінарного контенту, що поєднує глибинні моделі, графові структури знань і спеціалізовані конвеєри для опрацювання

відеоматеріалів. Особлива увага приділяється моделюванню станів інгредієнтів і розробці методів виявлення функціональних одиниць, які відображають зміст кулінарних дій у відеопослідовності. Результати роботи спрямовані на створення інтерпретованої, автоматизованої та стійкої до варіативності кулінарних даних системи аналізу, що має широкий спектр практичного застосування.

Актуальність роботи зумовлена стрімким зростанням обсягів кулінарного контенту, представленого у відеоплатформах, соціальних мережах, мобільних додатках та великих кулінарних базах даних. Для обробки такого контенту необхідні інтелектуальні системи, здатні автоматично розпізнавати інгредієнти, технологічні етапи, послідовності дій та характеристики готових страв. Сучасні підходи базуються переважно на глибинних моделях, але вони не забезпечують повноцінного розуміння процедурної логіки кулінарного процесу. Це створює потребу у розробці структурованих моделей знань, які дозволяють формалізувати кулінарні дії, стани об’єктів і причинно-наслідкові залежності між ними.

Актуальність підсилюється розвитком робототехніки, де автономні кухонні системи мають вміти інтерпретувати візуальні підказки, розпізнавати етапи приготування та точно відтворювати рецептурні процеси. Інтелектуальний аналіз кулінарного контенту також є ключовим для рекомендаційних сервісів, систем персоналізованого харчування, дієтологічних платформ і фуд-технологій. Окрім того, зростає потреба у методах оцінки калорійності на основі візуальних даних і автоматичного визначення складу страв.

З огляду на вказані тенденції розроблення комплексної методології інтелектуального аналізу кулінарних даних, що поєднує графові структури, комп’ютерний зір і багаторівневі конвеєри обробки відео, є вкрай актуальним і відповідає сучасним світовим науковим викликам.

Метою роботи є розробка та обґрунтування методології інтелектуального аналізу кулінарного контенту, що поєднує глибинні моделі,

графові структури знань та спеціалізовані алгоритми для розпізнавання об’єктів, станів і функціональних дій у кулінарному контенті.

Об’єктом дослідження є процеси інтелектуального аналізу кулінарного контенту.

Предметом дослідження є методи, моделі та інструменти інтелектуального аналізу і представлення знань, що забезпечують розпізнавання об’єктів, кулінарних станів, дій та процедурних залежностей у кулінарному контенті.

Завдання дослідження:

1. Проаналізувати предметну область інтелектуального аналізу кулінарного контенту та визначити ключові проблеми її автоматизації.

2. Дослідити методи глибинного навчання й графових структур для представлення кулінарних знань.

3. Розробити архітектуру функціональної об’єктно-орієнтованої мережі, що моделює кулінарні об’єкти, стани та дії.

4. Побудувати модель на основі ResNet для автоматичного визначення станів об’єктів.

5. Розробити методику автоматичного передбачення інгредієнтів багатоскладових страв на основі зображень.

Методи дослідження:

- методи обробки зображень;

- глибинні нейронні мережі (ResNet, Faster R-CNN та ін.);

- графові моделі знань і онтологічні структури;

- статистичний аналіз, методи оцінювання точності та матриць плутанини.

Наукова новизна

Запропоновано комплексну методологію інтелектуального аналізу кулінарного контенту, що поєднує графові структури знань і глибинні моделі. Запропоновано конвеєр розуміння кулінарного відео з інтеграцією розпізнавання функціональних одиниць і виведення графа завдань.

Розроблено модель передбачення інгредієнтів багатоскладових страв за їхнім зображенням, що підвищує семантичну інтерпретованість візуальних даних.

Практичне застосування результатів

Отримані результати можуть бути використані для створення інтелектуальних кулінарних асистентів, рекомендаційних систем, алгоритмів оцінки складу та калорійності страв, а також у системах автоматизації приготування їжі. Розроблена методика класифікації станів інгредієнтів може застосовуватися у додатках для контролю якості продуктів та у програмах харчового аналізу.

Структура магістерської роботи. Представлена робота складається зі вступу, трьох розділів та висновків. Загальний обсяг роботи становить 85 сторінок, і містить 32 рисунки, 12 таблиць, перелік використаних джерел із 42 позицій.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ІНТЕЛЕКТУАЛЬНОГО ВИДОБУВАННЯ ЗНАНЬ З КУЛІНАРНОГО КОНТЕНТУ

1.1. Автоматизація кулінарних процесів на основі візуалізації та структурування знань

Розвиток робототехнічного маніпулювання та його прикладних аспектів набув значного прискорення протягом останніх років. Сфери застосування та досліджень охоплюють такі галузі, як медицина, сільське господарство, промисловість, спорт, колаборативна робототехніка та побутові системи. Однією з актуальних і перспективних областей, що привернула увагу в контексті робототехніки, є кулінарна сфера, що охоплює як промислові, так і побутові сценарії.

Дана магістерська робота присвячена аналізу ключових проблем, пов'язаних із розробкою автоматизованих кулінарних робототехнічних систем, з акцентом на вимоги до сенсорного сприйняття середовища та алгоритмів маніпулювання.

Для забезпечення автономної діяльності робототехнічної системи в динамічному середовищі необхідний послідовний процес, який включає сенсорний огляд сцени (із застосуванням, наприклад, візуальних або глибинних камер), інтерпретацію отриманих даних за допомогою передових алгоритмів (зокрема, сучасних методів розуміння зображень та відео) та подальше адекватне маніпулювання.

Основна увага в цьому дослідженні зосереджена на проблемі візуального розуміння того, що сприймає робот, та на розробці методології для вилучення структурованих знань як із візуальної сцени, так і з будь-якого заданого кулінарного контенту. Ці знання є критично важливими для навчання роботів та автоматизованого розуміння сцени. З огляду на те, що сирі дані від робототехнічних датчиків зазвичай представлені у форматі потоків зображень (відео), необхідне використання алгоритмів для аналізу

зображень та відео з метою інтерпретації сцени. Оскільки пріоритетом є розробка автоматизованого кулінарного робота, ми зосереджуємося на розумінні сцен та відео, релевантних до кулінарного середовища.

1.1.1. Задачі візуального розуміння в кулінарній робототехніці

Для коректного аналізу сцени робот повинен спочатку ідентифікувати об'єкти (тобто інгредієнти), присутні у відеопотоці. Наприклад, ідентифікація помідора у відео свідчить про кулінарну активність, що вимагає використання цього інгредієнта.

1. Розпізнавання об'єктів (інгредієнтів). Однією з ключових проблем, що вирішуються в рамках цієї дисертації, є розпізнавання об'єктів (інгредієнтів). Для цього можуть бути застосовані алгоритми багатоміткової класифікації зображень та виявлення об'єктів.

2. Ідентифікація станів об'єктів. Окрім ідентифікації самого інгредієнта, критично важливим є визначення його стану. Розпізнавання станів сприяє як вилученню знань із відео, так і розумінню поточної сцени. Наприклад, якщо для приготування страви потрібен нарізаний помідор, а робот ідентифікує цілий помідор, він повинен ініціювати додаткову дію нарізання. Якщо ж помідор вже ідентифікований як нарізаний, підготовчі кроки не потрібні. У дисертації детально досліджується проблема візуальної ідентифікації станів об'єктів у кулінарному середовищі.

3. Оцінка порцій. Кількісні характеристики інгредієнтів (порції, наприклад, 1 склянка помідорів) є вирішальними для коректного виконання рецепта. Оцінка порцій може бути здійснена за допомогою візуальної оцінки та встановлення кореляцій між характеристиками інгредієнтів та їхніми візуальними ознаками.

4. Розпізнавання рухів (подій). Для повного вилучення знань із відео необхідно також розуміти кожну подію (тобто рух), що відбувається. Багато рухів (наприклад, нарізання) безпосередньо призводять до зміни стану об'єкта (наприклад, помідора), і їхнє взаємне визначення є корисним. Однак

деякі рухи (наприклад, наливання) можуть не мати безпосереднього фізичного впливу на стан інгредієнта, вимагаючи окремих кодувальних механізмів. У цій роботі також розробляється модель для розпізнавання руху у відео, демонструючи її важливість для ідентифікації кулінарних подій.

1.1.2. Структуроване представлення знань та графові моделі

Інгредієнти, їхні стани та порції не є незалежними сутностями; вони існують у взаємопов'язаних відношеннях. Наприклад, знаючи, що рецепт використовує борошно (навіть якщо воно візуально нерозпізнаване у готовому виробі, наприклад, пирозі), можна з високою ймовірністю передбачити наявність цукру.

Для ефективного моделювання цих сутностей необхідно враховувати їхні взаємні відношення та створювати асоціації через представлення знань або навчання моделі. Ми пропонуємо дослідити взаємозв'язки між об'єктами, діями та активностями та представити ці взаємозв'язки у формі графа. Граф використовується як структурована апріорна інформація для покращення розуміння відео.

Наприклад, відео, що демонструє приготування омлету, складається з послідовних дій. Кожна дія, як-от змішування яєць у мисці, задіює декілька об'єктів (миска, віночок, яйця). Для ідентифікації дії структурна інформація про зв'язки між об'єктами та рухами (змішування) є корисною.

Структурна інформація між послідовними діями також застосовується для інтерпретації активності. Наприклад, дія розбивання яєць у миску передує дії змішування яєць. Вбудовування цих інформативних структур у графічну апріорну структуру та використання цих вбудовувань для висновків на етапі тестування може суттєво покращити розуміння відео.

Головною метою цієї роботи є вилучення необхідних знань із візуального контенту (зокрема, кулінарного), включаючи інгредієнти, їхні стани, порції та рухи, асоційовані з активністю, що демонструється у відео. Для досягнення цієї мети застосовується комбінація відомих методів

представлення знань та алгоритмів машинного навчання з метою створення моделей, здатних кодувати знання про кулінарні відео та сцени маніпулювання. Вилучені знання використовуються для формування структурованих та послідовних даних у вигляді графа завдань, придатного для використання робототехнічною системою.

1.2. Огляд досліджень в області візуального сприйняття та представлення знань у кулінарній сфері

Представлений розділ містить огляд ключових досліджень, спрямованих на розробку алгоритмів для розуміння візуального контенту як у загальновживаних, так і в спеціалізованих кулінарних застосуваннях. Спершу ми розглянемо роботи, що використовують представлення знань для аналізу візуального контенту, а потім перейдемо до огляду методів комп'ютерного зору та обробки зображень, зокрема у контексті кулінарної робототехніки.

1.2.1. Представлення знань

Представлення знань (ПЗ) є фундаментальною доменною областю в робототехніці та штучному інтелекті, хоча його формальне визначення залишається гнучким. Початково концепція ПЗ виникла в галузі штучного інтелекту (ШІ) і стосувалася способу символічного представлення знань та їх подальшої автоматизованої обробки програмами міркування [1]. Це визначення акцентує увагу на логічних виразах, але не робить акцент на виведенні знань.

У контексті робототехніки розширюють це поняття, визначаючи ПЗ як "засіб представлення знань про дії робота та навколишнє середовище, а також пов'язування семантики цих концепцій з його власними внутрішніми компонентами для вирішення проблем через міркування та виведення". Таким чином, ПЗ містить інформацію про взаємозв'язки між об'єктами та

рухами, релевантними для робототехнічного застосування, що дозволяє роботу виконувати завдання та маніпуляції.

Представлення знань успішно застосовується в робототехніці та машинному навчанні, а також у обробці природної мови (ОПМ), зокрема у розробці Wordnet, Verbnet та Framenet [2].

WordNet є великою лексичною базою даних англійських лексем (іменників, дієслів, прикметників та прислівників), які організовані у набори когнітивних синонімів (синсети), кожен з яких виражає окреме поняття. Синсети пов'язані семантичними та лексичними зв'язками та використовуються в багатьох застосуваннях комп'ютерної лінгвістики [3].

VerbNet являє собою ієрархічну мережу англійських дієслів, що підтримує синтаксичні та семантичні зв'язки між ними [4].

В роботі [5] запропонували архітектуру на основі знань для вивчення мови з веб-текстів. Бази знань використовуються для відповідей на загальні запити, візуальні запити та запити, орієнтовані на кухню та інгредієнти, з використанням глибоких функцій. Даний підхід організований навколо спільної бази знань (БЗ), яка безперервно нарощується та використовується набором компонентів підсистеми навчання/читання, що реалізують комплементарні методи вилучення знань.

Стартова БЗ визначає онтологію (набір предикатів, що визначають категорії та відношення) та невелику кількість початкових прикладів (seed examples) для кожного предиката в цій онтології (наприклад, десяток прикладів міст). Мета підходу полягає в безперервному нарощуванні цієї БЗ шляхом читання та в навчанні кращому читанню.

Екземпляри категорій і відношень, додані до БЗ, поділяються на кандидатні факти (candidate facts) та переконання (beliefs). Компоненти підсистеми можуть читати з БЗ та звертатися до інших зовнішніх ресурсів (наприклад, текстових корпусів або Інтернету), а потім пропонувати нові кандидатні факти. Компоненти надають ймовірність для кожного запропонованого кандидата та резюме вихідного доказу, що його

підтверджує. Інтегратор Знань (ІЗ - Knowledge Integrator) аналізує ці запропоновані кандидатні факти та підвищує найбільш вагомо підтверджені з них до статусу переконань. Цей потік обробки зображено на рисунку 1.1.

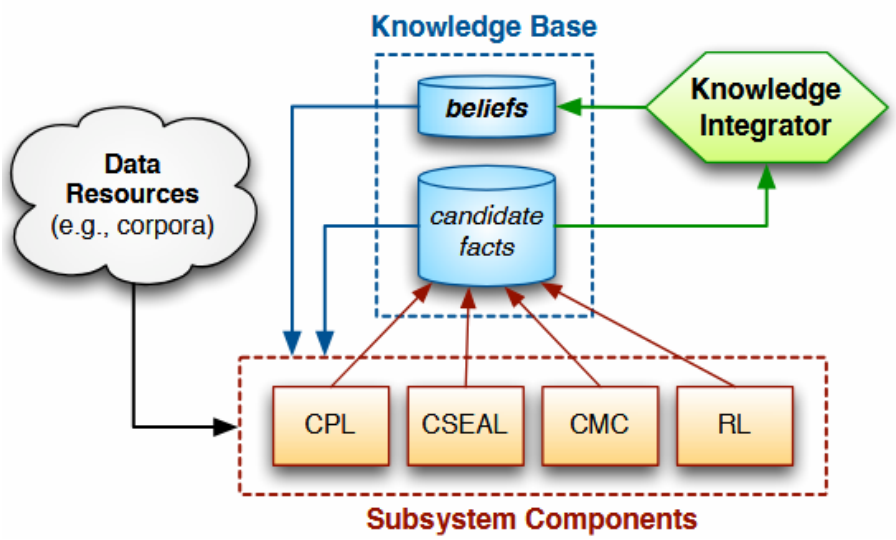


Рис. 1.1. Архітектура NELL - Never-Ending Language Learner

Методи на основі знань також ефективно застосовуються у візуальних застосуваннях:

- Онтологічна ієрархічна база знань для пошуку вмісту зображень та виявлення подій на відео.

- Розуміння сцени.

- Використання дескриптивної логіки для інтерпретації сцени.

- Комбінація різних представлень на основі знань із застосуванням методів машинного навчання та статистичних підходів.

- Візуальна структурована база знань для розпізнавання сцени та виявлення об'єктів [5].

Підхід з [5] схематично представлено на рис. 1.2. Тут використовується пошук зображень Google (Google Image Search) для завантаження тисяч зображень для кожної категорії об'єктів, сцен та атрибутів. Далі цей метод застосовує ітеративний підхід для очищення міток та напівавтоматичного навчання детекторів/класифікаторів.

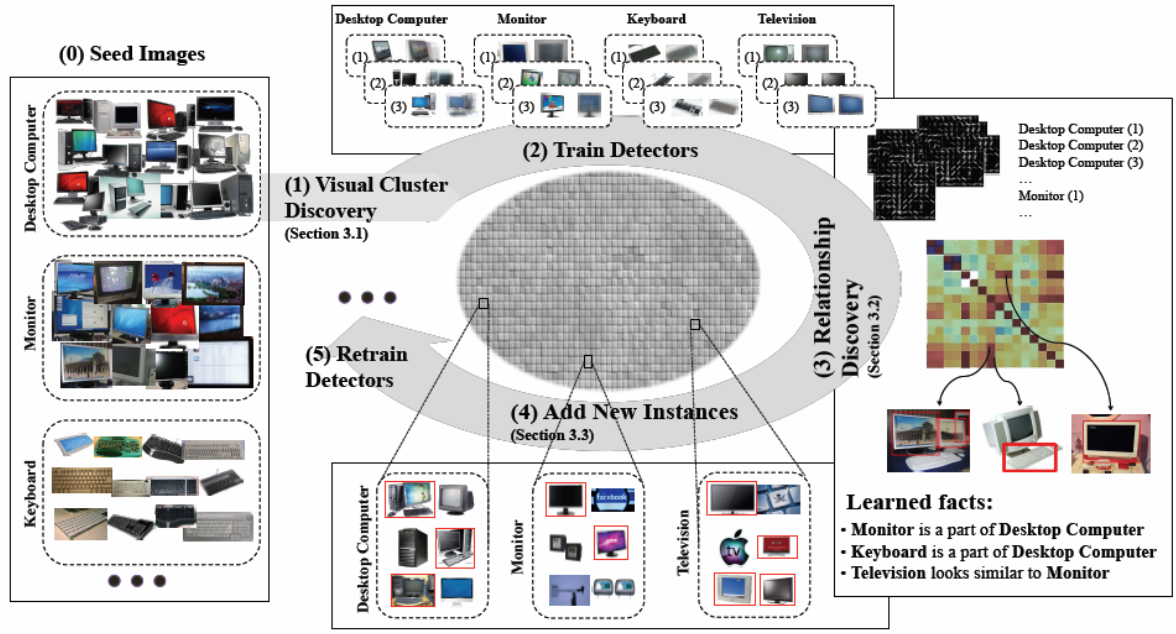


Рис. 1.2. Схема ітеративного підходу побудови структурованої бази знань для розпізнавання сцени та виявлення об'єктів

Для заданого концепту (наприклад, "автомобіль") спершу виявляють латентні візуальні підкатегорії та обмежувальні рамки для цих підкатегорій, використовуючи кластеризацію на основі екземплярів. Потім навчаються множинні детектори для концепту (по одному для кожної підкатегорії), використовуючи результати кластеризації та локалізації.

Ці детектори та класифікатори використовуються для виявлення об'єктів на мільйонах зображень з метою вивчення взаємозв'язків на основі статистики спільного виникнення (co-occurrence statistics). Тут застосовується той факт, що цікавить користувача - макробачення (macro-vision), і тому будується статистика спільного виникнення, використовуючи лише впевнені виявлення/класифікації.

Після встановлення взаємозв'язків, їх використовують разом із класифікаторами та детекторами для маркування великого набору "зашумлених" зображень. Зображення з найвищою впевненістю маркування додаються до пулу маркованих даних і використовуються для перенавчання моделей, після чого процес повторюється. На кожній ітерації навчаються кращі класифікатори та детектори, що, своєю чергою, допомагає вивчити

більше взаємозв'язків та додатково обмежити задачу напівавтоматичного навчання.

У роботі [6] проблема міркування про можливості об'єктів моделюється за допомогою представлення бази знань. Дослідження [7] пропонує візуальне представлення знань та набір даних для моделювання відносин на зображеннях. Зокрема, запропоновано метод на основі ПЗ для розпізнавання їжі на зображенні, що є близьким до нашого застосування. Відсутність структурованого представлення знань для спільного моделювання об'єктів та рухів стала мотивацією для застосування функціональної об'єктно-орієнтованої мережі для розуміння кулінарних відео.

1.2.2. Класифікація зображень та виявлення об'єктів

Зображення є важливими джерелами інформації. Останніми роками було розроблено численні алгоритми глибокого навчання для вилучення корисної інформації з зображень та виконання таких завдань, як класифікація зображень та виявлення об'єктів (наприклад, виявлення людей, автомобілів). У цьому підрозділі ми оглянемо роботи, що можуть слугувати базовими алгоритмами для вилучення знань із кулінарного контенту.

У завданнях класифікації зображень модель приймає на вхід одне зображення та генерує на виході одну мітку класу. Припущення в цих моделях полягає у наявності одного домінуючого об'єкта (інгредієнта) на зображенні.

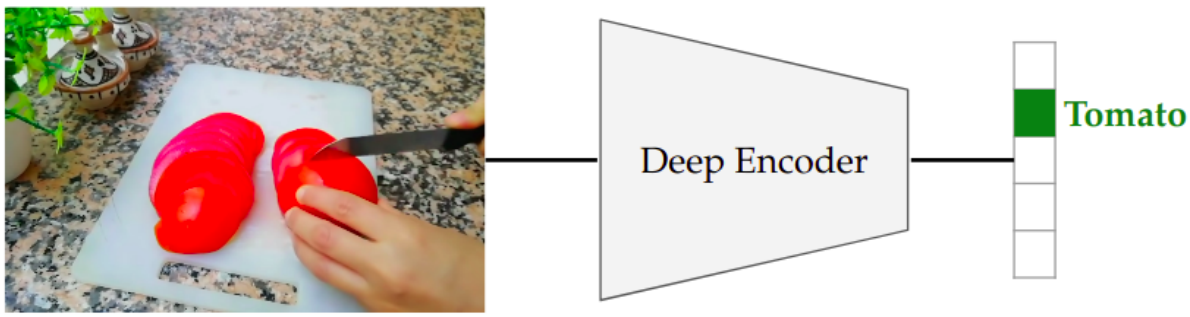


Рис. 1.3. Ілюстрація простої моделі класифікації зображень

Модель на рис. 1.3 приймає на вхід зображення та повертає на вихід вектор впевненості (ймовірностей) для кожного класу

Більшість поширених моделей класифікації зображень використовують підходи глибокого навчання. Глибокі згорткові мережі (CNN) домінують у цій галузі з 2012 року після публікації Alexnet. Пізніше були запропоновані глибші архітектури, такі як VGG [7], які здатні захоплювати глибші та багатші ознаки із зображення, забезпечуючи високу продуктивність.

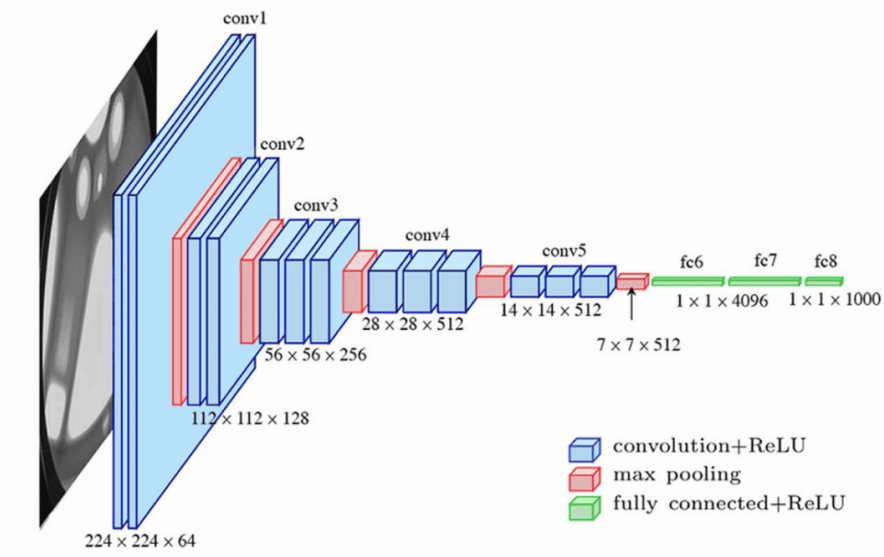


Рис. 1.4. Архітектура VGG моделі мережі

Інші моделі, включаючи Inception та Resnet [8], запропонували оптимізації для глибоких CNN, такі як врахування різних розмірів фільтрів у шарі та покращене поширення градієнта у надглибоких моделях. У цій роботі ми використовуємо варіанти Resnet для навчання, тонкого налаштування або вилучення ознак із зображень.

У застосуваннях багатоміткової класифікації зображень передбачається, що модель приймає одне зображення та генерує кілька вихідних міток. Іншими словами, ці моделі здатні спільно моделювати різні концепції в одному зображенні одночасно.

Деякі підходи пропонують поєднання CNN з рекурентними нейронними мережами (RNN) [9]. CNN вилучає семантичні ознаки із

зображення, тоді як RNN моделює асоціації між зображенням і мітками, а також між самими мітками, передбачаючи впорядкований взаємозв'язок. У [10] запропоновано модель для максимізації точності підмножин за допомогою RNN. Дослідження [11] пропонує канонічний корельований автоенкодер (C2AE) для багатоміткової класифікації, виконуючи спільне вбудовування ознак та міток і вводячи функцію втрат, чутливу до кореляції міток, для відновлення передбачених вихідних міток.

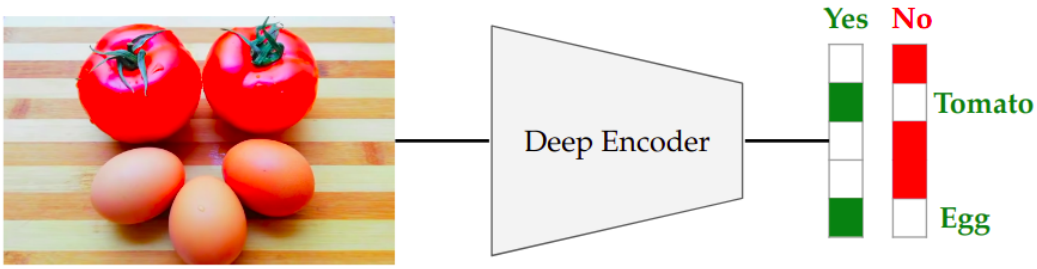


Рис. 1.5. Ілюстрація простої моделі багатоміткової класифікації зображень

Модель на рис. 1.5 приймає на вхід зображення та повертає на вихід бінарну впевненість класифікації для кожного імені класу.

Моделі підписування зображень (Image Captioning) також належать до категорії багатоміткової класифікації. Вони спочатку створюють словник слів та призначають кожному слову ідентифікатори [12]. Ідентифікатори перетворюються на one-hot вбудовування і використовуються для передбачення моделі [13]. На відміну від спільної генерації міток у багатомітковій класифікації, у моделях підписування мітки генеруються покроково та вимагають впорядкованої залежності між згенерованими виходами. Ці моделі часто використовують авторегресивні моделі, такі як RNN (наприклад, LSTM) [14], або механізми на основі трансформерів та уваги. Зокрема, [15] пропонує модель передбачення інгредієнтів на основі трансформера, яка демонструє вищу продуктивність порівняно з існуючими багатомітковими класифікаторами з функцією втрат розподілу цільових значень.

У застосуваннях виявлення об'єктів модель приймає одне зображення та генерує кілька обмежувальних рамок з відповідними мітками класів [16]. Сучасні методи виявлення об'єктів, подібно до класифікації, використовують глибокі згорткові мережі як основу для вилучення ознак та одночасної локалізації та маркування об'єктів.

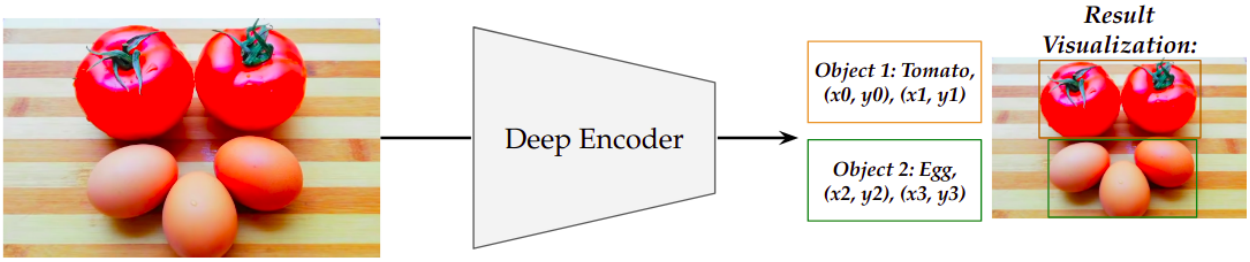


Рис. 1.6. Ілюстрація моделі виявлення об'єктів, яка приймає на вхід зображення та повертає обмежувальні рамки та їхні назви класів

Faster R-CNN - цей двостадійний детектор спочатку обробляє зображення через ConvNet, отримуючи карти ознак. Потім мережа пропозицій регіонів (RPN) застосовується до карт ознак для генерування пропозицій об'єктів. Ці пропозиції нормалізуються за розміром і класифікуються, що призводить до списку обмежувальних рамок та їхніх міток.

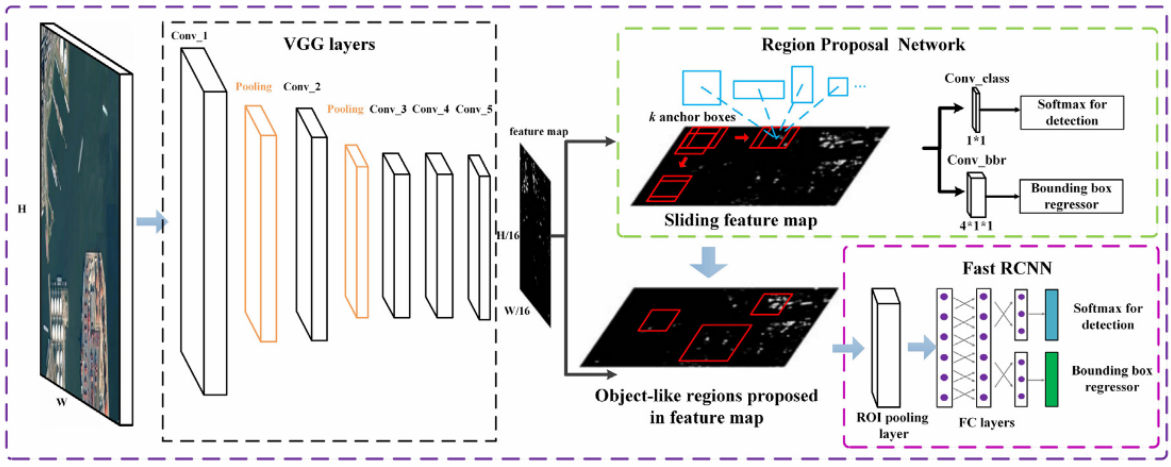


Рис. 1.7. Архітектура Faster R-CNN

Single Shot MultiBox Detector (SSD) [16] - це одностадійний детектор який запускає ConvNet для вхідного зображення, обчислює карту ознак, а потім застосовує невелику згорткову мережу (3x3) для передбачення обмежувальних рамок. SSD використовує анкерні рамки з різними співвідношеннями сторін і передбачає зміщення до рамки. Для забезпечення масштабної інваріантності, SSD передбачає обмежувальні рамки після кількох згорткових шарів, кожен з яких функціонує на різному масштабі.

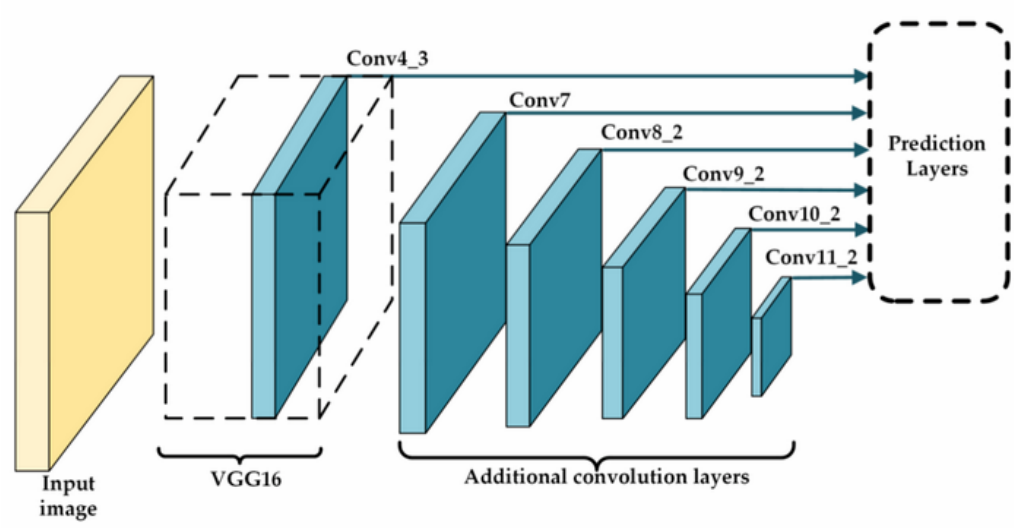


Рис. 1.8. Модель Single-Shot MultiBox Detector (SSD), що складається з архітектури VGG та додаткових згорткових шарів

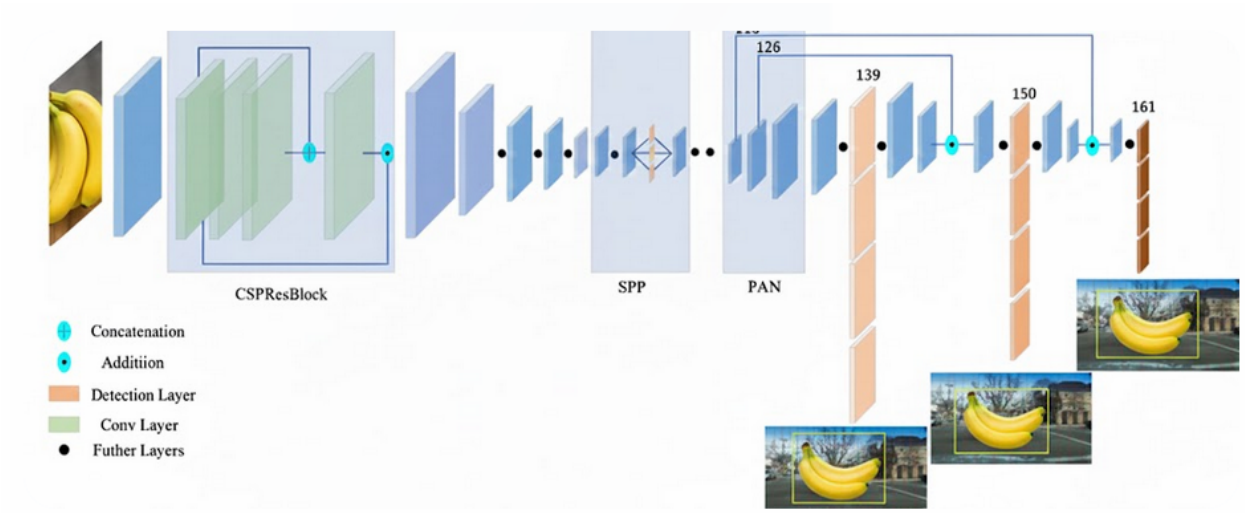


Рис. 1.9. YOLO модель

You Only Look Once (YOLO) [17] формулює виявлення об'єктів як задачу регресії, використовуючи уніфіковане виявлення як одностадійну мережу. У ньому зображення ділиться на сітку, і кожен квадрат сітки передбачає потенційні обмежувальні рамки та оцінки впевненості. Єдина нейронна мережа виконує операції на повних зображеннях за одну оцінку. Це уніфікований підхід робить YOLO надзвичайно швидким (базова модель – 45 к/с, Fast YOLO – 155 к/с). Хоча перша версія YOLO може мати більше помилок локалізації, вона менш схильна до хибних спрацьовувань на фоні порівняно з іншими детекторами. Крім того, YOLO навчається дуже загальних представлень об'єктів, добре узагальнюючи на інші домени (наприклад, твори мистецтва).

Retina-net [18] - це ще одна одностадійна мережа виявлення об'єктів, яка на момент її представлення демонструвала вищу продуктивність, ніж усі двостадійні мережі (наприклад, Faster R-CNN), при швидкості, співмірній з одностадійними детекторами. У Retinanet було запропоновано новий тип втрат – Focal loss, який ефективно вирішує проблему дисбалансу класів у задачі виявлення об'єктів. Focal loss фокусує оптимізацію на складніших зразках, запобігаючи домінуванню легких зразків.

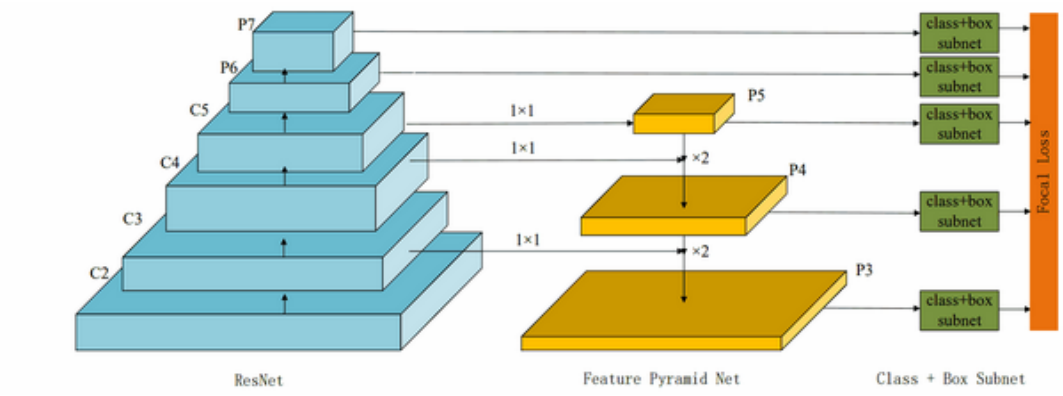


Рис. 1.10. Архітектура Retina-net

Незважаючи на те, що YOLO, SSD та Retinanet є швидшими алгоритмами, Faster R-CNN може досягати кращої точності при використанні більш просунутих CNN-основ.

1.3. Аналіз методів розуміння відео та кулінарних станів

1.3.1. Аналіз необробленого відео

Галузь розуміння відео охоплює широкий спектр досліджень. Деякі роботи покладаються на дороговартісні конфігурації, такі як фізичні датчики або додаткові модальності (наприклад, текстові дані) [19]. Інші дослідження зосереджуються на комплексному аналізі просторово-часового контенту відеопослідовності для маркування дій [20] або використовують просторово-часові ознаки людини (наприклад, моделі суглобів або пози) для класифікації дій [21]. Однак ці методи часто демонструють обмежену стійкість до варіацій ракурсу, збільшення масштабу та оклюзії.

Одночасна сегментація та розуміння відео [22] також є поширеною дослідницькою областю, хоча ці підходи зазвичай не враховують об'єкти або варіації в позі. Деякі методи витягують та аналізують вибірку кадрів для узагальнення подій відео [23] та швидкого виявлення й концентрації аномалій [24]. В [22] запропонували метод, який вбудовує структуру у глибоку модель для інтеграції знань із глибокими моделями з метою розпізнавання активності.

Існує також різноманіття багатовидових застосувань, особливо в системах відеоспостереження. Інформація з кількох камер може підвищити узагальнення подій або розуміння завдань. Було запропоновано методи для обробки багатокамерних сценаріїв, такі як:

- Узагальнення подій у багатовидових відео за допомогою глибокого навчання.

- Виявлення та узагальнення події у багатовидових відеоспостереженнях із застосуванням бустингу.

- Використання ансамблевого методу машинного навчання [23].

Хоча аспекти багатовидового розуміння відео не є основним фокусом даної роботи, запропонована архітектура має потенціал для розгортання в багатовидових системах.

1.3.2. Представлення знань для розуміння відео

Було запропоновано різноманітні підходи для використання представлення знань (ПЗ) у контексті розуміння відео:

- Використання семантично-візуальних баз знань, таких як FrameNet та ImageNet, для моделювання багатих подієво-центричних концепцій та їхніх відношень з метою виявлення подій на відео [24].

- Застосування структури знань та ймовірнісної структури для розпізнавання активності.

- Використання семантичних представлень для виявлення подій [25].

В роботі [26] використовують об'єкти, дії та їхні зв'язки у вигляді графів та застосовують імітоване відпалювання для виведення подій з урахуванням часових зв'язків. Раніше запропонували баєсівську структуру, яка використовує рухи об'єктів та їхні відношення для підвищення надійності розпізнавання об'єктів. Ця модель також дозволяє роботам вивчати інтерактивні функції об'єктів із людських демонстрацій [27].

Аналіз об'єктної інформації є ключовим аспектом для розпізнавання активності. Метод у [28] використовує просторові та функціональні обмеження на відношення між об'єктами та рухами для семантичної інтерпретації відео. Інші приклади робіт із розуміння відео з використанням інформації про об'єкти включають:

- Моделювання взаємного контексту пози людини та об'єктів за допомогою моделі випадкового поля.

- Моделювання відношень між частинами об'єктів та людьми на сцені за допомогою контекстних дескрипторів сцени та баєсівського навчання.

- Кодування об'єктів для класифікації та локалізації дій.

Усі ці підходи припускають, що дія виконується людиною, тому поза людини є важливою для їхньої методології. Ми наслідуємо шлях включення об'єктів, розширюючи його до мети розпізнавання дій та виведення активності, використовуючи раніше запропоновану мережу представлення знань. Наша робота відрізняється від зазначених методів розпізнавання

активності на основі об'єктів тим, що наші відео не містять людини та її пози. Ми використовуємо лише людську руку та її розташування (за наявності на сцені) як ознаки для інтерпретації відео.

1.3.2 Розуміння кулінарних станів

Наскільки відомо, спеціалізовані дослідження, присвячені лише ідентифікації станів об'єктів на зображенні, не проводилися. У цьому розділі ми обговорюємо роботи з класифікації зображень, підписування зображень та розуміння, які є релевантними або послугували джерелом натхнення для нашого дослідження. Сучасна класифікація зображень домінує на основі згорткових нейронних мереж (CNN). Першу еволюційну глибоку модель для класифікації зображень представили в [26]. Після цього були представлені більш глибокі та вдосконалені архітектури, такі як VGG, Googlenet та Resnet. Покращення шляхом комбінування цих мереж також були представлені в [27]. Усі ці роботи сфокусовані на класифікації об'єктів на зображеннях, не враховуючи стани об'єктів.

У [28] автори продемонстрували важливість використання частин об'єктів у розпізнаванні дії на зображенні, моделюючи дії людини на основі частин та атрибутів. Це слугує доказом того, як частини об'єктів та стани можуть сприяти розпізнаванню об'єкта або розумінню зображення.

Роботи, такі як [29] та [30], займаються створенням підписів до зображень або відео. Наприклад, [29] використовує атрибути та їхню взаємодію з глибокими мережами для генерації підписів. Інші дослідження, як-от [30] виконують багатоміткову класифікацію на одному зображенні за допомогою глибоких архітектур на основі RNN та CNN. Хоча ці статті надають множинні мітки для зображення, вони не розглядають стани об'єктів як окрему мітку для зображення.

Спільною рисою цих робіт є аналіз зображення з метою його розуміння. Проблема ідентифікації стану, натхненна цим аспектом, також спрямована на поглиблення розуміння зображень.

Проведено низку робіт у сфері кулінарних зображень та відео:

- Системи розпізнавання їжі для дієтичного аналізу.
- Розпізнавання фруктів.
- Розпізнавання інгредієнтів для пошуку рецептів [31].

Це приклади досліджень, які зосереджені на розпізнаванні або виявленні інгредієнтів на зображенні. Деякі роботи, виконують розпізнавання їжі на відео для розуміння всього відео та асоціації його з рецептом або дією. Інші статті, такі як [32], фокусуються на розпізнаванні активності з кулінарних відео. Хоча ці роботи сприяють розумінню кулінарних зображень та відео, жодна з них явно не зосереджується на станах. Наскільки нам відомо, наше дослідження є першим, що безпосередньо розглядає цю проблему в контексті кулінарних зображень або відео.

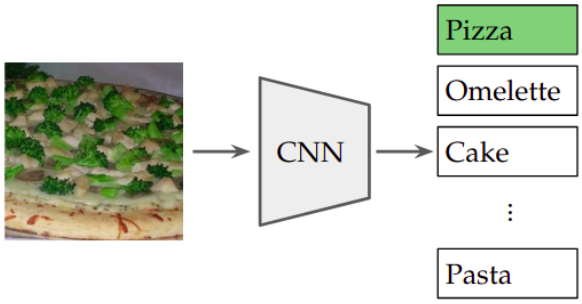
1.4. Спеціалізовані застосування комп'ютерного зору у кулінарній сфері

1.4.1. Класифікація страв

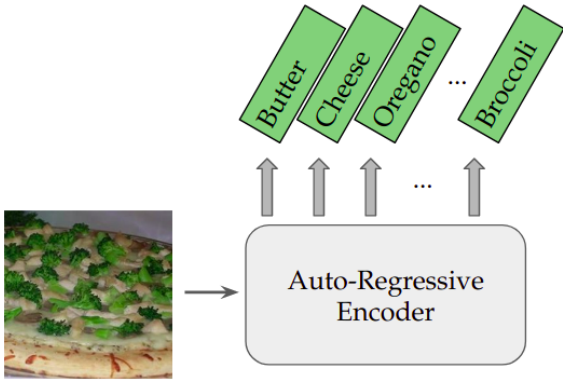
Класифікація страв (або їжі) може розглядатися як вузькоспеціалізоване застосування загальної задачі класифікації зображень. Існують роботи, які надають експериментальні дослідження з розпізнавання типів їжі (або страв) за одним заданим зображенням на невеликих наборах даних [33].

Багато застосувань класифікації страв інтегрують невізуальний контекст, такий як геолокація, для підвищення точності класифікації [34]. Для вирішення динамічної та мінливої природи класифікації їжі запропонували модель на основі персоналізації. Спільною рисою запропонованих робіт у цій галузі є обмежений масштаб набору даних та використання моделі глибокого навчання для вирішення завдання. У даній

роботі ми створюємо власний набір типів страв і використовуємо сучасну мережу глибокого навчання для виконання цього завдання.



а) Класифікація страв



б) Генерація інгредієнтів

Рис. 1.11. Ілюстрація моделі класифікації страв (повертає одну мітку на зображення) та моделі генерації інгредієнтів (повертає множинні мітки на зображення).

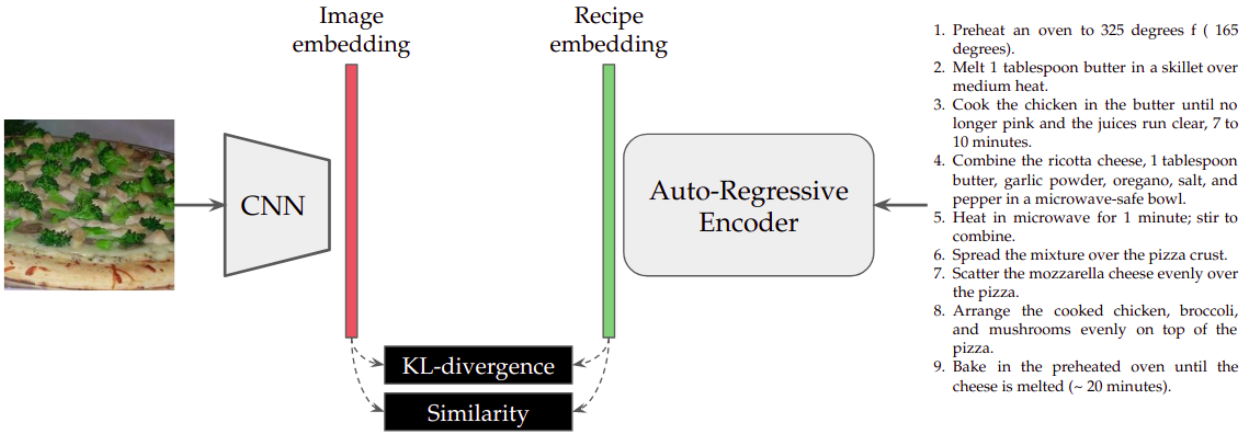
1.4.2. Розпізнавання інгредієнтів та рецептів

Дослідження у сфері розпізнавання інгредієнтів можна умовно поділити на дві основні категорії: на основі пошуку (retrieval-based) та на основі передбачення (prediction-based).

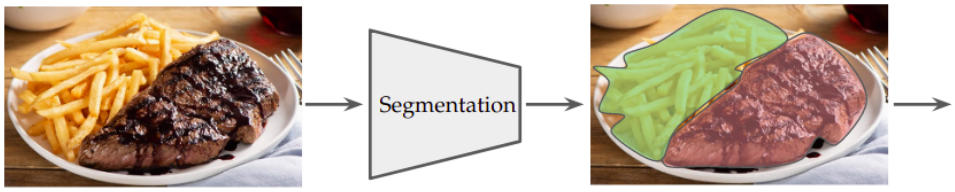
Підходи на основі пошуку. У цих застосуваннях список інгредієнтів або весь рецепт витягується шляхом створення вбудовування (embedding) зображення та отримання відповідного результату з набору даних. Цей

корпус робіт вимагає, щоб передбачувана комбінація інгредієнтів належала до фіксованого набору, наявного в одному з наборів даних.

Підходи на основі передбачення. Для подолання проблеми фіксованого набору виникли методи передбачення інгредієнтів, натхненні багатокласовим моделюванням, рекурентним підписуванням зображень та авторегресивними методами передбачення списків.



а) Пошук рецептів



	Fries	Steak
grams	200	220
calorie	630	620
total	1250	

б) Оцінка порцій та калорійності

Рис. 1.12. Приклади моделей для застосувань у кулінарній сфері: моделі для пошуку рецептів та оцінки порцій і калорійності

Розпізнавання станів інгредієнтів є іншою, недостатньо вивченою областю. Введення нових наборів даних про стани інгредієнтів або формулювання проблеми станів як задачі класифікації зображень чи багатокласового маркування (наприклад, кортеж "інгредієнт-стан") [35] є прикладами досліджень у цій галузі.

Однією з останніх робіт, яку ми використовуємо як базову для передбачення інгредієнтів, є дослідження "зворотного приготування їжі". У цій роботі інгредієнти та рецепти генеруються авторегресивним способом за допомогою моделі трансформера.

1.4.3. Оцінка порцій та калорійності

Дослідження порцій інгредієнтів проводилися переважно в контексті оцінки калорійності. У більшості робіт [36] порції інгредієнтів (наприклад, яблуко) визначаються після сегментації зображення та обчислення розміру за допомогою різних підходів (наприклад, геометрія, 3D-моделювання). Це використовується для оцінки калорійності на дуже простих кулінарних зображеннях у невеликих наборах даних.

Крім того, підхід до порцій з точки зору візуального розпізнавання та сегментації може бути нежиттєздатним у стравах, де інгредієнт не візуально розрізняється (наприклад, курятина у супі). Тому ми підходимо до проблеми порцій за допомогою методу, заснованого на самоувазі та запитах (self-attention and query-based approach), використовуючи великомасштабний набір даних (тобто Recipe1M).

Оцінка калорійності зображення привернула значну увагу. Деякі методи пропонують багатостадійні конвеєри для передбачення категорій/інгредієнтів їжі, визначення порцій/розмірів та оцінки калорійності. Деякі з цих методів вимагають двох вхідних зображень для визначення глибини та сегментації їжі. Ці алгоритми використовують модельні або глибокі методи для етапу розпізнавання та стандартні таблиці харчової цінності для оцінки калорійності [37].

Інші літературні джерела безпосередньо надають оцінки калорійності зображення їжі [38], передбачаючи категорію їжі та прямо відображаючи її на споживання калорій, з або без довідкового об'єкта. Недоліком цих методів є те, що вони не враховують різноманітність інгредієнтів, які можуть бути присутніми у різних версіях однієї страви. Деякі роботи пропонують метод

CNN на основі прямого зображення, який враховує кілька видів їжі на одному зображенні, але не включає інгредієнти, що містяться у стравах, для оцінки калорійності. Крім того, у [38] автори запропонували баєсівську структуру оцінки харчового балансу, яка враховує обмежену кількість категорій їжі з обмеженою кількістю класів, кожен з яких має обмежену кількість дискретних значень.

Більшість робіт, виконаних у галузі оцінки калорійності, припускають, що зображення стосуються їжі з чітко сегментованими межами і не розглядають більш складну їжу, таку як змішані або приготовані страви, де інгредієнти є нечіткими. Іншою проблемою є те, що використовуваний набір даних є дуже малим і має низьку різноманітність, а зображення зроблені у добре контрольованому середовищі. З іншого боку, існують літературні джерела, які використовують інгредієнти для оцінки калорійності на основі зображень. У [39] запропоновано метод на основі глибокого навчання для одночасного вивчення калорій, категорій, інгредієнтів та кулінарних інструкцій. Набори даних, такі як японський набір фотографій їжі з анотаціями калорійності та американський набір фотографій їжі з анотаціями калорійності, містять відповідні анотації.

Висновки до розділу

У першому розділі було проведено всебічне дослідження предметної області інтелектуального видобування знань із кулінарного контенту, що дозволило систематизувати сучасні наукові підходи до аналізу зображень, відео та структурованих даних у кулінарній сфері. Особливу увагу приділено ролі візуального розуміння при автоматизації кулінарних процесів, де комп’ютерний зір є ключовим елементом для розпізнавання інгредієнтів, кулінарних інструментів та дій. Було встановлено, що графові моделі та онтології істотно підсилюють можливість інтерпретації візуальної інформації, оскільки дозволяють формалізувати складні міжоб’єктні

взаємозв’язки. Аналіз існуючих методів класифікації зображень і виявлення об’єктів довів їхню ефективність, але також виявив обмеження, пов’язані з неоднозначністю візуальних даних у реальних кулінарних умовах. Дослідження методів розуміння відео показало, що задачі сегментації дій, побудови послідовностей та відстеження станів є значно складнішими, ніж статична класифікація об’єктів. Аналіз спеціалізованих застосувань штучного інтелекту у кулінарній сфері засвідчив високу перспективність автоматизації класифікації страв, оцінки їхньої калорійності та розпізнавання компонентів. У цілому розділ сформував фундамент для подальшої розробки методології інтелектуального аналізу кулінарного контенту, обґрунтувавши потребу у комплексних підходах.

РОЗДІЛ 2. МЕТОДОЛОГІЇ ТА РІШЕННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КУЛІНАРНОГО КОНТЕНТУ

2.1. Графові структури як апріорні знання для розуміння кулінарного відео

Ключовим викликом для автоматизованого робототехнічного маніпулювання є розуміння сцени. Для ефективного розуміння сцени робототехнічні системи повинні здійснити візуальний аналіз та інтерпретацію сутностей з метою подальшого прийняття рішень. Розуміння відео виступає основним загальним рішенням цієї проблеми.

Розуміння відео є складною мультидисциплінарною темою, що вимагає успішного виконання кількох етапів, кожен з яких є самостійною та активною дослідницькою галуззю. Типово, це включає:

- Автоматичне розбиття відео на атомарні дії.

- Успішне розпізнавання активностей та об'єктів у межах атомарного відеокліпу.

- Виведення значущого розуміння на основі ідентифікованих активностей та об'єктів.

Хоча для кожного етапу (розпізнавання об'єктів, розпізнавання активностей, розбиття відео) проводиться інтенсивне навчання, ці процеси традиційно виконуються ізольовано.

2.1.1. Структурований підхід на основі графових моделей

Ми пропонуємо підхід, заснований на моделюванні взаємозв'язків між об'єктами, діями та активностями та їх представленні у вигляді графа. Цей граф використовується як структурована апріорна інформація для підвищення ефективності розуміння відео.

Наприклад, відео, що демонструє приготування омлету шеф-кухарем, складається з послідовних дій. Кожна дія, як-от змішування яєць у мисці,

задіює множинні об'єкти (наприклад, миска, віночок, яйця). Для ідентифікації дій структурна інформація про зв'язки між об'єктами та рухами (змішування) є корисною. Зокрема, якщо відомо, що яйця можна змішувати за допомогою віночка, встановлюється асоціація між об'єктом "віночок" та об'єктами "яйце" і "миска".

Структурна інформація між послідовними діями також може бути застосована для інтерпретації активності на відео. Наприклад, оскільки дія розбивання яєць у миску передує змішуванню яєць, це знання дозволяє передбачити поточну дію "змішування" на основі ідентифікації попередньої дії. Вбудовування цих інформативних структур у попередню графічну структуру та використання цих вбудовувань для виведення на етапі тестування може покращити розуміння відео.

2.1.2. Використання функціональної об'єктно-орієнтованої мережі

Для розпізнавання дій (наприклад, "перемішування яєць") ми використовуємо координацію, закодовану у вузлах об'єктів (наприклад, "миска" або "яйця") та вузлах руху (наприклад, "перемішування") мережі на основі знань. Мережа, яка використовується для виведення завдань, називається функціональною об'єктно-орієнтованою мережею (FOON), і вона кодує знання про послідовність дій, що відбуваються одна за одною. Використовуючи цю мережу, ми представляємо потужний об'єктно-орієнтований алгоритм виведення для розпізнавання дій та активностей.

Функціональна об'єктно-орієнтована мережа є двочастковою мережею, що містить два типи вузлів: вузли об'єктів і вузли руху. Вузли об'єктів ідентифікуються за їхнім типом об'єкта, їхнім спостережуваним станом, а також, якщо це контейнер, його вмістом. Рухи ідентифікуються за типом. Ця графічна структура є аналогічною мережам Петрі, де вузли об'єктів паралельні вузлам місць, а вузли руху — вузлам переходів. Крім того, FOON є орієнтованим, ациклічним графом, що означає, що в ньому існують ребра, які вказують на потік або послідовність зміни стану об'єкта.

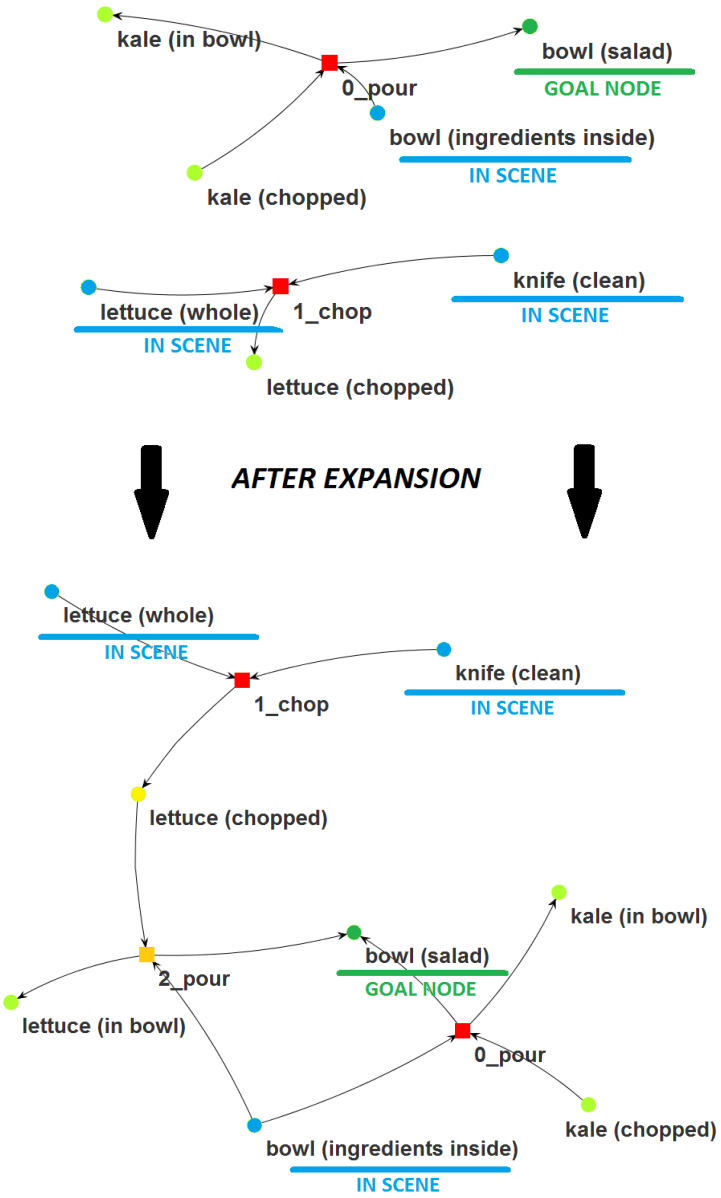


Рис. 2.1. Приклад того, як розширення (expansion) допомагає додавати знання, які можуть бути корисними для вирішення ситуаційної проблеми.

У цьому прикладі (рис. 2.1) ми бажаємо приготувати салат (вузол цілі, позначений темно-зеленим кольором), використовуючи салат-латук та інші інгредієнти, наявні в середовищі (позначені синім кольором). Однак початково ми володіємо знаннями лише про приготування салатів із капустою кейл. Використовуючи подібність (similarity), ми можемо пов'язати знання про нарізання латуку та його додавання до миски з іншими інгредієнтами для приготування салату.

Ми пропонуємо конвеєр, який використовує локалізацію об'єктів та їхні функції руху для ідентифікації активних об'єктів у межах дії. Додатково, ми навчаємо глибоку модель для загального розпізнавання руху, що є критично важливим у випадках, коли об'єкт (наприклад, сіль у руці шеф-кухаря) не піддається легкому виявленню.

Виявлені об'єкти та рух подаються на етап виведення з FOON, який надає список кандидатів функціональних одиниць, що можуть бути асоційовані з поточною дією (наприклад, "розбивання яйця в миску"). Послідовні передбачені функціональні одиниці оцінюються для розуміння загальної активності, що виконується на відео (наприклад, "приготування омлету").

2.2. Архітектура функціональної об'єктно-орієнтованої мережі для інтелектуального кодування знань про кулінарні завдання

Функціональна об'єктно-орієнтована мережа (FOON) — це представлення знань, розроблене для кодування знань про завдання маніпулювання та, як наслідок, можливості об'єктів. FOON може бути використана роботом для вирішення завдань маніпулювання за заданою цільовою метою. Наразі основна увага FOON зосереджена на вивченні активностей у кулінарній та кухонній галузях, але її архітектура дозволяє розширення на інші предметні області та середовища.

FOON є спрямованим ациклічним графом (Directed Acyclic Graph, DAG), який містить два типи вузлів: об'єкт та рух, що робить його двопартійною мережею (рис. 2.2).

Вузли об'єктів (NO) визначаються як сутності, якими маніпулює або на які впливає демонстратор. Вузол об'єкта ідентифікується за його типом об'єкта, станом об'єкта та ідентифікатором руху, який позначає, чи рухається об'єкт під час активності. Об'єкти також можуть функціонувати як

контейнери для інших об'єктів, а кожен вузол може бути описаний списком інгредієнтів.

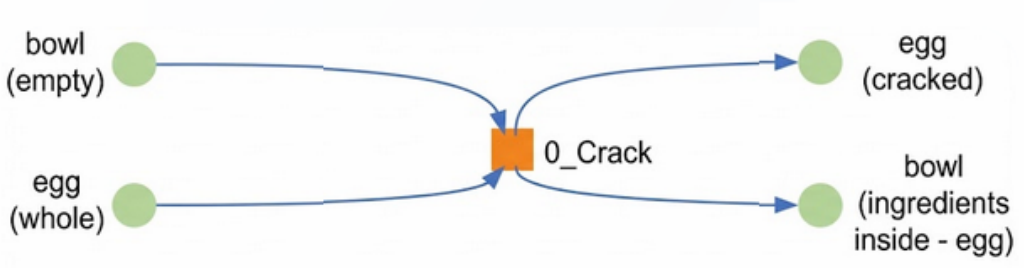


Рис. 2.2. Ілюстрація функціональної одиниці із вхідними та вихідними вузлами об'єктів, їхніми станами та вузлом руху

Вузли руху описують дію, що застосовується до об'єктів (наприклад, різання або змішування). Вузли руху ідентифікуються лише за їхнім типом руху.

У цій графічній структурі, подібно до звичайних двопартійностей, ребра з'єднують пару вузлів, а саме вузол об'єкт-рух. Напрямок ребра вказує на послідовність, у якій об'єкт може змінювати свій стан через рух, подібно до мереж Петрі (Petri Nets), які вимагають активації переходів для запуску вузлів місця.

FOON складається з підкомпонентів, або навчальних одиниць, які називаються функціональними одиницями (Functional Units, FUs). Кожна функціональна одиниця описує одну атомарну дію, що спостерігається в активності (активність або підграф можна розглядати як серію дій). Наприклад, в активності приготування яєчні, одна функціональна одиниця може описувати дію розбивання яйця, а інша — дію змішування яєць у мисці.

Функціональна одиниця описує перехід станів об'єктів до та після маніпуляційного руху:

- Вхідні вузли об'єктів: Об'єкти до маніпуляції.

- Вихідні вузли об'єктів: Об'єкти після маніпуляції.

Наша увага в цій роботі зосереджена на генерації цих функціональних одиниць безпосередньо з навчальних відео. Універсальною FOON

називається сукупність підграфів (активностей), об'єднаних для комбінування знань та видалення дублікатів.

Кожна функціональна одиниця має три ключові компоненти:

- Вхідні вузли об'єктів.

- Вихідні вузли об'єктів.

- Вузол руху, що описує дію, яка потенційно викликає зміну стану вхідних об'єктів (зміна стану не завжди відбувається).

Кожна функціональна одиниця також описується часовими кадрами, у яких вона спостерігається в активності.

2.2.1. Конструкція функціональної мережі

Граф, представлений на рисунку 2.3 , складається з вузлів, отриманих з 65 відео, які були анотовані у вигляді підграфів. Ці підграфи складаються з функціональних одиниць, що відображають кожен окремий крок у кулінарній процедурі.

Ребра проводяться між парою вузлів об'єкт-рух, де вузли об'єктів — це ті, що спостерігаються в дії, а вузол руху описує дію. Під час створення цих підграфів було складено список об'єктів та рухів для забезпечення узгодженості міток (оскільки підграфи створювалися кількома добровольцями).

При додаванні нової інформації (підграфів) з інших наборів даних, необхідна лише їхня анотація для відповідності формату наших графів та аналіз для підтвердження коректності міток. Процедура злиття додає ці нові проаналізовані функціональні одиниці до мережі, щоб запобігти дублюванню. Ця процедура злиття детальніше описана в нашій попередній роботі.

Саме в цьому контексті наша запропонована робота набуває актуальності: автоматична генерація підграфів з відео (особливо з інших наборів даних) є складним завданням, а ручне анотування є трудомістким процесом.

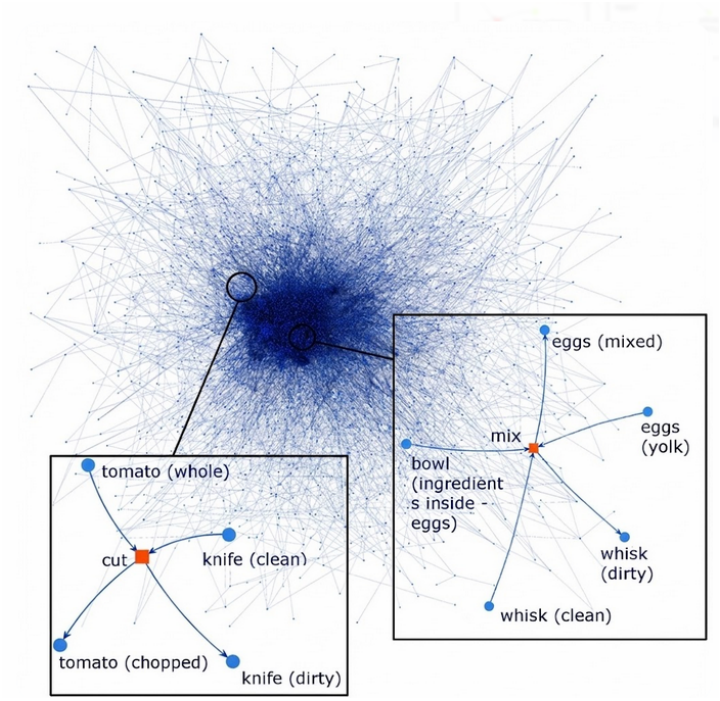


Рис. 2.3. Ілюстрація універсальної FOON, що складається з 4955 вузлів (вузлів об'єктів та рухів)

Ідеально, FOON вивчається безпосередньо з людських демонстрацій (відео чи спостережень) та генерується автоматично. Проте, на ранніх етапах створення, ми вирішили вручну маркувати відео з YouTube як підграфи. У майбутньому планується розширення FOON.

Після запису всіх функціональних одиниць для відео, підграф аналізувався для забезпечення узгодженості всіх міток об'єктів та рухів з усіма іншими підграфами. Кожен підграф потім об'єднувався в єдину мережу, названу універсальною FOON (GFOON). Процедура злиття полягає у порівнянні кожної функціональної одиниці в усіх підграфах зі списком одиниць у GFOON та додаванні тих одиниць, яких немає.

Загалом універсальна мережа містить:

- 1853 вузли об'єктів
- 3102 вузли руху
- 15656 ребер

На рисунку 2.3 зображено мережу, описану цими статистичними даними.

2.2.2. Порівняння функціональної мережі та інших представлень знань

FOON не є першим представленням знань, що стосується розуміння відео. Основна відмінність FOON від попередніх робіт полягає в тому, що вони не розглядають спільне представлення як об'єктів, так і рухів.

Наша робота побудована на основі теорії можливостей (affordance theory), а наступні дослідження підтверджують зв'язок між маніпуляціями та об'єктами. Наша мета — створити графічне представлення маніпуляцій, де об'єкти та рухи описують можливості.

Щодо графічних представлень, попередні роботи захоплюють знання за допомогою ймовірнісних графічних методів або семантичних графів/дерев. Однак вони не створюють базу знань активності з демонстрацій, яку потім можна було б використовувати для виконання (можливих) нових маніпуляцій. Дослідження можливостей здебільшого зосереджувалися на моделюванні відносин між об'єктами та простими діями для передбачення їхнього ефекту.

Більш загальною формою представлення, подібною до FOON, є мережі Петрі, де:

- Вузли місця схожі на вузли об'єктів.

- Вузли переходів схожі на вузли руху.

Для активації або виконання вузла переходу необхідні певні вхідні вузли місця, аналогічно тому, як вхідні вузли об'єктів повинні бути доступними для виконання заданого руху маніпуляції.

2.3. Чотириетапний конвеєр розуміння кулінарного відео на основі функціональної мережі

Ми пропонуємо чотириетапний конвеєр для розуміння відео, спрямований на автоматичне призначення структурованих знань візуальному контенту. Конвеєр ідентифікує функціональні об'єкти та рухи у відеопослідовності (що відповідають атомарній дії) і використовує ці дані

разом із представленням знань (FOON) для призначення функціональній одиниці (Functional Unit, FU) мітки для події, що відбувається. Дія стосується окремої атомарної події, тоді як послідовність дій представляє всю активність (рецепт). Послідовно ідентифіковані дії в подальшому аналізуються цілісно для розуміння активності, що виконується на відео.

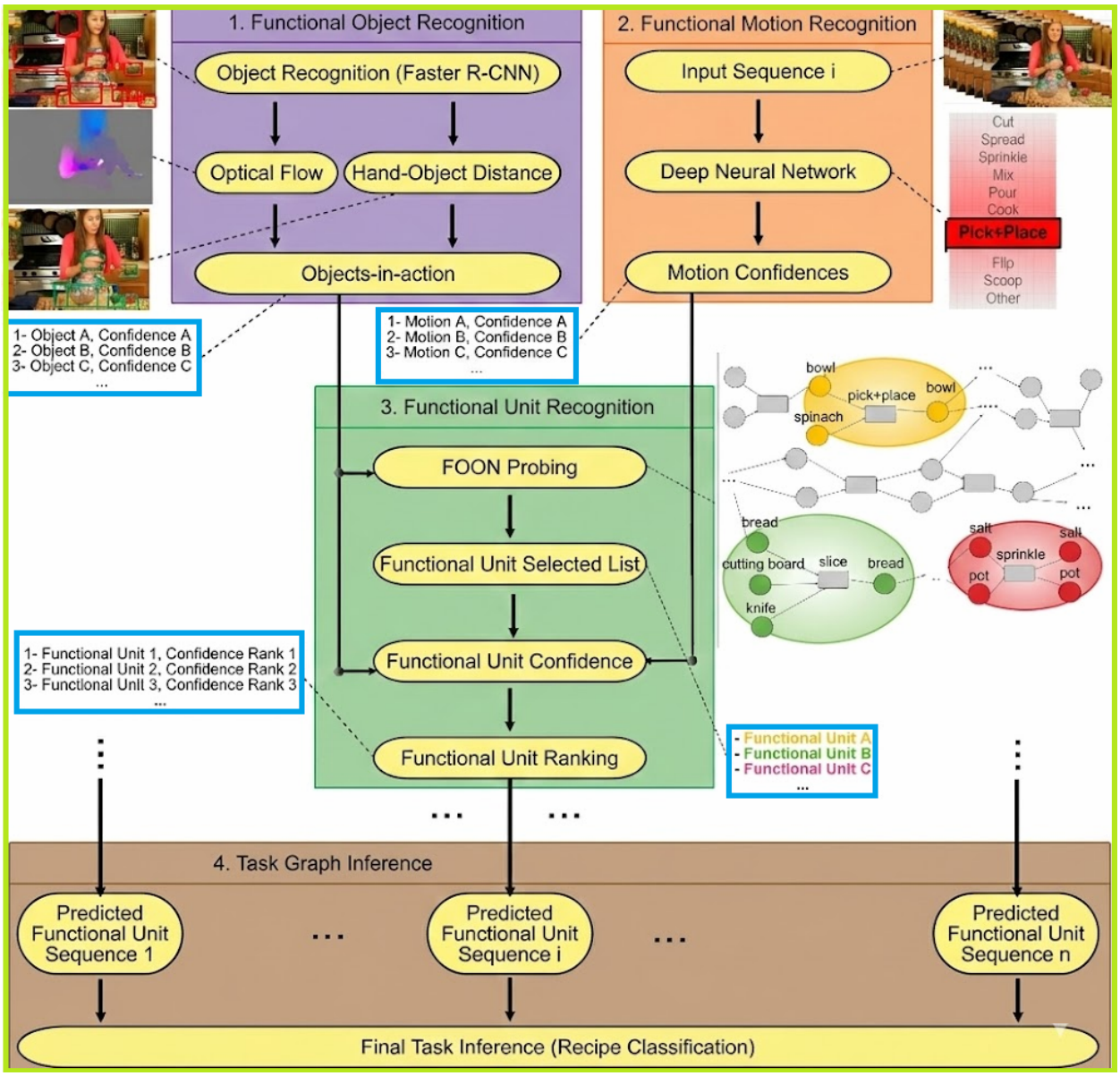


Рис. 2.4. Чотириетапний конвеєр для автоматичного розуміння кулінарного відео

Етапи конвеєра:

- Розпізнавання функціональних об'єктів.

- Розпізнавання функціональних рухів.

- Розпізнавання функціональних одиниць.

- Виведення графа завдань.

На першому етапі, розпізнавання функціональних об'єктів, ідентифікуються всі об'єкти, і їм призначаються оцінки корисності на сцені. На другому етапі, розпізнавання функціональних рухів, кожна дія (розбитий відеокліп) класифікується у відповідний клас руху. Використовуючи результати перших двох етапів та їхні відповідності FOON, кожна дія аналізується та асоціюється з функціональною одиницею на етапі розпізнавання функціональних одиниць. Кожній розпізнаній дії з одного відео призначається функціональна одиниця, яка потім шукається в графі FOON, щоб зрештою бути класифікованою як цілісна активність (рецепт). Цей останній етап називається виведенням графа завдань.

2.3.1. Розпізнавання функціональних об'єктів

Ми застосовуємо алгоритм Faster R-CNN для локалізації та маркування об'єктів на сцені. Faster R-CNN є двокомпонентною згортковою мережею, що одночасно виявляє пропозиції об'єктів та виконує їхню класифікацію. Виходом мережі є набір обмежувальних рамок та відповідні мітки класів об'єктів.

Далі ми ідентифікуємо об'єкти, які фактично використовуються у відеопослідовності, які ми називаємо об'єктами-в-дії (objects-in-action), використовуючи три метрики:

- Близькість людської руки до об'єкта.

- Величина оптичного потоку.

- Частота, з якою об'єкти спостерігалися у відео.

2.3.2. Розпізнавання функціональних рухів

У деяких сценаріях FOON не може однозначно ідентифікувати дію лише на основі ознак об'єкта. Наприклад, знання про те, що об'єкти "миска"

та "яйце" є об'єктами-в-дії, може призвести до кількох виведень FOON, оскільки різні функціональні одиниці містять ці об'єкти, але мають різні вузли руху (наприклад, "наливання" або "розбивання"). В іншому прикладі, при посипанні сіллю рукою, візуально важко розрізнити об'єкт "сіль", але рух руки вказує на дію "посипання".

Для вирішення цих проблем ми виконуємо тонке налаштування глибокої мережі CNN+LSTM, змінюючи її останній шар на 10 класів.

Архітектурно мережа складається з частини CNN та частини LSTM. Кадри відеопослідовності послідовно подаються на вхід CNN, вихід якого подається на шар LSTM. Виходи шару LSTM усереднюються для отримання остаточного передбачення класу руху.

Архітектура CNN містить п'ять згорткових шарів та два повністю з'єднані шари. Перші п'ять згорткових шарів і один повністю з'єднаний шар подаються на одношаровий рекурентний шар LSTM. За виходом LSTM слідує шар класифікації.

Ми модифікували останній шар, щоб він містив десять нейронів, які відповідають 10 обраним нами типам руху. Ми навчаємо архітектури CNN та CNN+LSTM окремо, використовуючи навчені ваги з [45] і виконуючи навчання лише для останнього шару класифікації. У звіті наводяться кращі результати архітектури CNN+LSTM.

Кожен клас руху в цій глибокій архітектурі пов'язаний з набором вузлів руху в FOON. Мережа призначає оцінки впевненості кожному класу руху, що відображає ймовірність того, що цей клас буде призначений як мітка дії. Вихід із цієї глибокої мережі використовується для розрахунку впевненості для кожної кандидатської функціональної одиниці на наступному етапі.

2.3.3. Розпізнавання функціональних одиниць та виведення графа завдань

Ми визначаємо семантичне значення відео шляхом асоціації дій з функціональними одиницями (FUs). Об'єкти-в-дії використовуються для

пошуку в універсальній FOON з метою ідентифікації кандидатських функціональних одиниць.

Ці кандидатські FU оцінюються на основі консолідованої оцінки впевненості, яка інтегрує:

- Впевненість об'єктів, отриману на етапі розпізнавання функціональних об'єктів.

- Впевненість руху, отриману на етапі розпізнавання функціональних рухів.

Ця оцінка кількісно визначає, наскільки кожна кандидатська FU пов'язана з поточною дією. Список кандидатських FU сортується за впевненістю, і одиниці з найвищою впевненістю асоціюються з поточною дією.

Для ідентифікації активності (послідовності дій) на відео, ідентифіковані дії протягом усього відео використовуються разом із пошуком FOON для передбачення найбільш ймовірної мітки активності для цього відео.

2.4. Розпізнавання функціональних об'єктів у відео за допомогою алгоритму Faster R-CNN

Ми здійснюємо розпізнавання та локалізацію всіх об'єктів у відеопослідовності (що відповідає атомарній дії) за допомогою алгоритму Faster R-CNN. Faster R-CNN (Faster Region-based Convolutional Neural Network) — це сучасний двостадійний алгоритм глибокого навчання, призначений для виявлення об'єктів (object detection) на зображеннях. Його ключова інновація полягає в інтеграції етапу генерації пропозицій регіонів у саму нейронну мережу, що значно підвищує швидкість порівняно з попередніми моделями R-CNN та Fast R-CNN.

Після локалізації об'єктів ми кількісно оцінюємо ступінь залученості (involvement) кожного об'єкта в поточну дію. Ця оцінка відбувається шляхом

вилучення ознак оптичного потоку та обчислення відстаней між рукою та об'єктом у кожному кадрі послідовності. Результатом є формування списку найбільш релевантних об'єктів, який називається об'єкти-в-дії (objects-in-action).

На цьому етапі конвеєра ми використовуємо обмежувальну рамку, асоційовану з кожним об'єктом. Об'єкти, що рідше зустрічаються у відео, виключаються. Для обчислення середньої відстані об'єкта від руки використовується центральна точка обмежувальних рамок, отриманих за допомогою Faster R-CNN. Отримані значення відстаней нормалізуються за допомогою гаусового розподілу.

Також обчислюється оптичний потік об'єктів у межах відеопослідовності. Оцінений оптичний потік та позиції об'єктів інтегруються для оцінки потоку кожного об'єкта. Об'єктам із вищою величиною оптичного потоку призначається вища оцінка впевненості, оскільки вище значення вказує на вищу ймовірність руху об'єкта, а отже, на вищу ймовірність його використання у відеопослідовності.

Інтеграція цих метрик для оцінки впевненості об'єкта (confObject) здійснюється за допомогою лінійної комбінації, як показано у наступному рівнянні:

$$\mathrm{conf}_{\mathrm{Object}} = \alpha \cdot c_{\mathrm{flow}} + \beta \cdot c_{\mathrm{dist}} + \gamma \cdot c_{\mathrm{freq}}$$

де:

c_{flow} — впевненість оптичного потоку.

c_{dist} — впевненість відстані до руки.

c_{freq} — впевненість частоти спостереження об'єкта.

α, β, γ — коефіцієнти, які налаштовуються вручну і відображають відносну вагу кожного фактора у кінцевій оцінці впевненості.

На рисунку 2.5 зображено процедуру ідентифікації об'єктів-в-дії для простої дії збивання яєць, використовуючи три метрики, згадані в рівнянні. У даному прикладі об'єкти "яйце", "віночок" та "миска" отримують найвищі

оцінки впевненості як об'єкти-в-дії, тоді як об'єкти "сковорода" та "плита" мають нижчу впевненість.

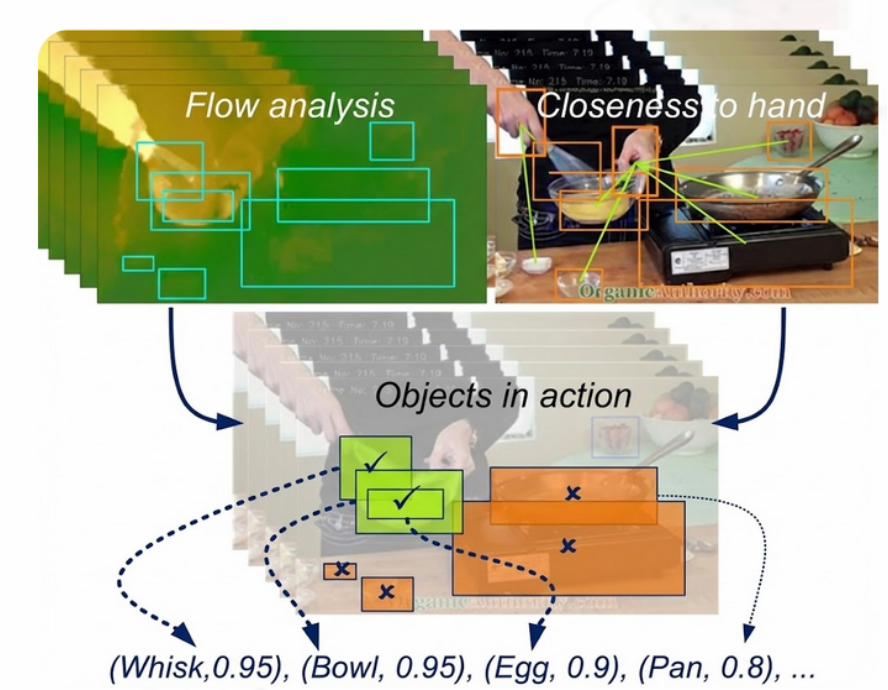


Рис. 2.5. Приклад, що демонструє процедуру ідентифікації об'єктів-в-дії (objects-in-action)

2.5. Процес розпізнавання функціональних одиниць та оцінка впевненості

Кожна дія, що відбувається у відео, асоціюється з найбільш релевантною функціональною одиницею (FU) з FOON. Для забезпечення коректності асоціації непов'язані функціональні одиниці підлягають фільтрації. Фільтрація здійснюється на основі оцінки впевненості функціональної одиниці та зондування (probing), як обговорюється нижче.

Конвеєр спочатку надає список об'єктів-в-дії (objects-in-action), які використовуються в поточній дії. Ці об'єкти використовуються для пошуку у FOON та ідентифікації функціональних одиниць, що їх містять. Виявлені

функціональні одиниці пропонуються як кандидатські функціональні одиниці (Candidate FUs) для поточної дії.

Кожна кандидатська FU містить вузли об'єктів (NO), які можуть бути або не бути включені до списку об'єктів-в-дії.

Використаний Набір (Nused): Перетин між вузлами об'єктів FU та об'єктами-в-дії.

Невикористаний Набір (Nnotused): Решта вузлів об'єктів.

Ці два набори використовуються для визначення того, чи слід підтримувати (bonus) або штрафувати (penalty) кандидатську FU. Впевненість кандидатської функціональної одиниці (confFOON) оцінюється за наступним рівнянням:

$$\text{conf}_{\text{FOON}} = \frac{\sum_{n=1}^{N_{\text{used}}} \text{conf}_n}{N_{\text{used}}} - \text{penalty} + \kappa \cdot \text{bonus}$$

У цьому рівнянні:

N_{used} — кількість вузлів об'єктів у використаному наборі.

$conf_n$ — впевненість кожного з цих об'єктів.

bonus — оцінюється на основі піксельного перекриття (pixel overlap) всіх об'єктів, що використовуються в кандидатській FU, представляючи ступінь взаємодії між об'єктами.

penalty — розраховується за наступним рівнянням і представляє штраф за розбіжності між об'єктами FU та об'єктами-в-дії.

$$\text{penalty} = \sum_{m=1}^{N_{\text{notused}}} \lambda \cdot \text{conf}_m + \sum_{k=1}^{N_{\text{extra}}} \eta \cdot \text{conf}_k$$

Термін штрафу формується на основі:

- Впевненості об'єктів ($conf_m$), які є в списку об'єктів-в-дії, але не використовуються в кандидатській FU (невикористаний набір, Nnotused).

- Впевненості об'єктів ($conf_k$), які не перераховані як об'єкти-в-дії, але використовуються в кандидатській FU (N_{extra}).

Константи κ,λ,η налаштовуються для балансування впливу бонусу та двох складових штрафу.

Рисунок 2.6 ілюструє процедуру оцінки впевненості для кандидатської функціональної одиниці. Алгоритм для розрахунку впевненості представлено в лістингу 2.1.

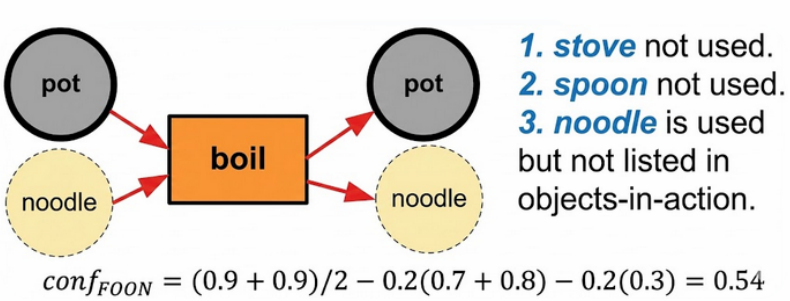


Рис. 2.6. Ілюстрація оцінки впевненості функціональної одиниці.

У цьому прикладі ідентифікованими об'єктами-в-дії є каструля, ложка та плита з впевненістю 0.9, 0.8 та 0.7 відповідно (λ=η=0.2).

Лістинг 2.1. Алгоритм розрахунку впевненості

```
list = ∅ // містить об'єкти та їх впевненості
for object ∈ sequence do
    c_dist = abs(object - hand)
    c_freq = frequency(object)
    c_flow = opticalFlow(object)
    conf_object = α · c_flow + β · c_dist + γ · c_freq
    list.append((object, conf_object))
end for
list.sort()
topK = list.selectTopK() // об'єкти в дії
// Знайти всі кандидатські функціональні одиниці, пов'язані з топ-K об'єктами
candidates = FOONLookUp(topK)
for c ∈ candidates do
    nodes = c.getObjects()
    overlap = objectOverlap(nodes.objects, topK.objects)
    N_used = size(overlap)
    bonus = pixelOverlap(nodes.objects)
    unused = (topK - overlap) + (nodes - overlap)
    penalty = average(unused.confidences)
    conf_FOON(c) = (Σ conf_object(n) for n in N_used) / N_used - penalty + κ · bonus
end for
```

Впевненість, обчислена у попередньому рівнянні, зосереджується виключно на взаємодії об'єктів. Для включення функціонального руху ми інтегруємо виходи навченої глибокої архітектури CNN+LSTM для класифікації руху. Вихід мережі CNN+LSTM (10 оцінок впевненості) ранжується.

Остаточна впевненість функціональних одиниць ($conf_{motion}$) об'єднує впевненість, оцінену на основі об'єктної взаємодії ($conf_{FOON}$), з результатами мережі CNN+LSTM ($conf_{LSTM}$), як показано у наступному рівнянні:

$$\mathrm{conf}_{\mathrm{motion}} = \mathrm{conf}_{\mathrm{FOON}} + \alpha \cdot \mathrm{conf}_{\mathrm{LSTM}}$$

де коефіцієнт α балансує вплив кожного з цих параметрів.

На етапі зондування кожен об'єкт окремо шукається у FOON. Ідентифікується список кандидатських функціональних одиниць, що містять цей об'єкт. Об'єкти з нижчою впевненістю (confobject) виключаються зі списку, щоб зменшити кількість потенційних об'єктів-в-дії та, як наслідок, кількість зондованих об'єктів і кандидатських FU.

Кожна зондована FU містить вузли об'єктів, які можуть бути або не бути спостережені в поточній дії. Використовується значення перетину (intersection) між об'єктами, включеними в зондовану FU, та ідентифікованими об'єктами у відеопослідовності. Кандидатські FU зі значенням перетину, меншим за певний поріг, виключаються. Для решти FU обчислюються значення впевненості, і ті, що мають найвищі значення, вибираються як найбільш ймовірні FU, пов'язані з поточною дією.

2.6. Комплексна оцінка конвеєра розуміння відео від розпізнавання дій до класифікації рецептів

Для аналізу ефективності функціональної об'єктно-орієнтованої мережі (FOON) у контексті розуміння відео було проведено низку експериментів.

Набір даних. Універсальна FOON була сформована на основі 338 навчальних відео, які включали анотовані демонстрації, використані для створення FOON, а також відео з набору даних MPII Cooking Activities. Загальний обсяг FOON на момент проведення досліджень становив понад 3000 функціональні одиниці (ФО). Для проведення поточних експериментів окрема підмножина відеопослідовностей була додатково анотована вручну обмежувальними рамками об'єктів та їхніми категоріями.

Як тестовий набір було використано 11 кулінарних відео із загальної вибірки, що містили 55 функціональних одиниць. Через обмежену кількість навчальних даних для розпізнавання об'єктів, брак точних анотацій (ground truth) та невелику кількість екземплярів деяких типів ФО в наборі даних, ми застосували процедуру перехресної валідації "leave-one-out" (залишити один): в кожній із 11 ітерацій одне відео повністю виключалося, а решта 337 відео використовувалися для побудови FOON, специфічної для даної ітерації.

Було реалізовано три основні експерименти, що оцінюють конвеєр на основі як ручного, так і автоматичного маркування об'єктів:

- Порівняння розпізнавання ФО з використанням виключно пошуку FOON проти злиття FOON із розпізнаванням руху.

- Оцінка розуміння відео (розпізнавання ФО) з FOON та без її використання.

- Виведення завдань (класифікація рецептів).

2.6.1. Метрика оцінювання

Результати оцінювалися за метрикою перекриття (overlap) між кандидатською функціональною одиницею та відповідною їй справжньою ФО. Ця метрика обчислювалася для кожної атомарної дії окремо.

Перекриття обчислювалося лише за умови, що вузол руху кандидатської ФО був еквівалентним вузлу руху справжньої ФО. В іншому випадку, точність (Precision) та повнота (Recall) вважалися нульовими.

Точність та повнота обчислювалися на основі перетину об'єктних вузлів:

- точність - перекриття, поділене на кількість об'єктних вузлів у кандидатській ФО.

- повнота - перекриття, поділене на кількість об'єктних вузлів у справжній ФО.

Для аналізу відео були розбиті на складові атомарні дії за допомогою міток часу, наданих в універсальній FOON.

Кожна послідовність дій подавалася в алгоритм, який ідентифікував найкращу ФО на основі метрик впевненості. Точність та повнота обчислювалися для 55 ФО тестового набору в діапазоні до топ-10 результатів, як показано на рисунку 2.7.

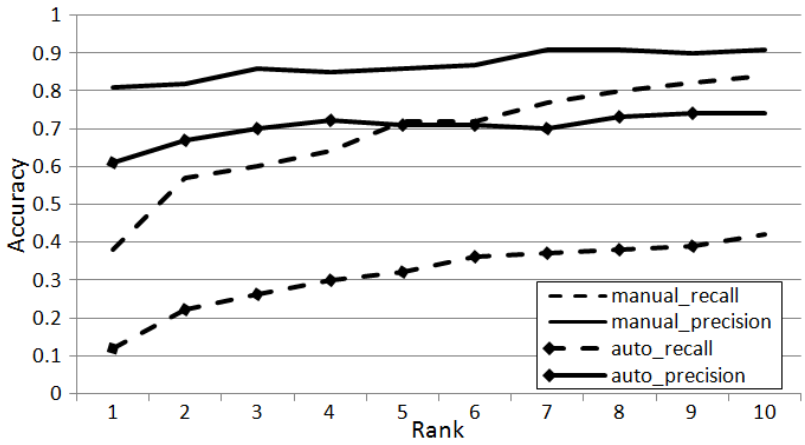


Рис. 2.7. Точність та повнота, отримані при ручному та автоматичному розпізнаванні об'єктів для топ-10 результатів

Аналіз рисунка 2.7 демонструє, що точність постійно перевищує 80%. Це свідчить про високу достовірність, коли алгоритм припускає, що об'єкт використовується у ФО. Однак, ймовірно, мали місце випадки пропуску об'єктів на відео, що відображається у менших значеннях повноти.

Для підвищення якості розпізнавання руху було інтегровано розпізнавання руху на основі тонко налаштованої глибокої мережі CNN+LSTM. Було визначено 10 класів руху (дев'ять найбільш поширених типів руху з FOON та клас "інші").

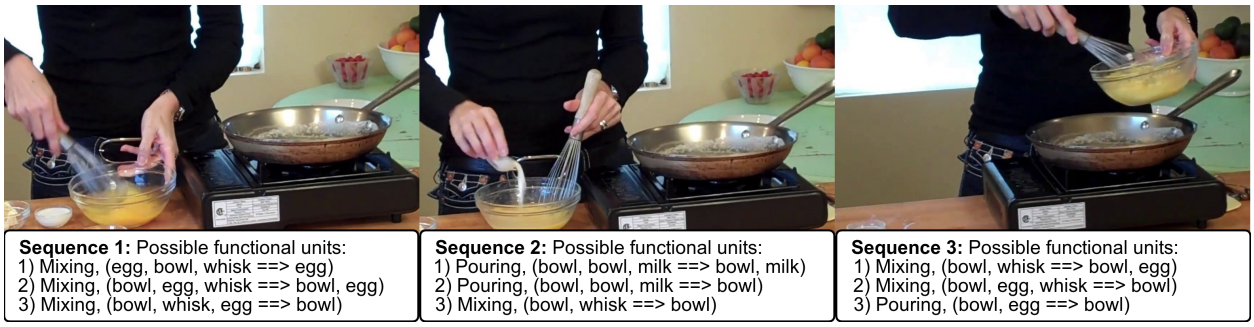


Рис. 2.8. Приклад розпізнавання функціональної одиниці з використанням розмічених та вручну розбитих послідовностей для рецепту яєчні

Мережа обробляла послідовності RGB та оптичного потоку, а її виходи (10 значень впевненості) використовувалися для обчислення остаточної впевненості кандидатських ФО за рівнянням наведеним в попередньому розділі). Для запобігання погіршенню результатів через обмежену точність глибокої мережі руху (~47%), коефіцієнт α у рівнянні був встановлений на рівні менше 0.2.

Передбачення вважалося правильним, якщо:

- Вузол руху ідентифікованої ФО був еквівалентним вузлу руху справжньої ФО (з урахуванням синонімів, як-от "збивання" і "змішування").

- Перекриття об'єктних вузлів становило більше 80%.

Таблиця 2.1 демонструє, що точність розпізнавання ФО значно зросла після об'єднання розпізнавання руху з пошуком FOON. Це підтверджує, що інтеграція автоматичного розпізнавання руху покращує ідентифікацію вузлів руху та, відповідно, розпізнавання ФО.

Таблиця 2.1

Точність передбачення для розпізнавання функціональних одиниць з використанням FOON та розпізнавання руху

	Using FOON	Using FOON + Motion Recognition
Top 1	56%	64%
Top 3	75%	84%
Top 5	80%	89%
Top 10	89%	98%

Порівняння окремих компонентів показало, що автоматичне розпізнавання руху досягло 67% точності (для 10 класів), тоді як розпізнавання ФО без руху досягло 61% точності (для більше ніж 50 типів руху). Хоча ці результати безпосередньо не порівнювані, вони підтверджують, що рух є цінною ознакою для об'єднання з FOON.

Перекриття між об'єктами-в-дії та ідентифікованими ФО становило 84%. Нижча загальна точність розпізнавання ФО порівняно з цим значенням пояснюється помилками в ідентифікації вузлів руху.

Рис. 2.9. Знімки подій, їхнє істинне представлення функціональної одиниці та передбачена функціональна одиниця

В іншому експерименті ми застосували конвеєр у поєднанні з розпізнаванням руху для автоматично розпізнаних об'єктів та представили його результати в топ-10 на рисунку 2.7. Хоча розпізнавання об'єктів є важливим етапом конвеєра, який можна вдосконалити, ми не розглядаємо його детальніше, оскільки це не є нашою основною метою в цій роботі.

Знімки різних послідовностей із їхнім істинним представленням (ground truth) та ідентифікованими функціональними одиницями зображені на рисунку 2.9

Оцінка конвеєра з точки зору розуміння відео здійснювалася за допомогою метрики перекриття, усередненої для всіх дій у кожному відео.

Як показано на рисунку 2.10, конвеєр демонструє здатність до розуміння відео, особливо у межах топ-5 результатів. Нижчі значення повноти можуть бути пов'язані з помилками на етапі ідентифікації об'єктів-в-дії.

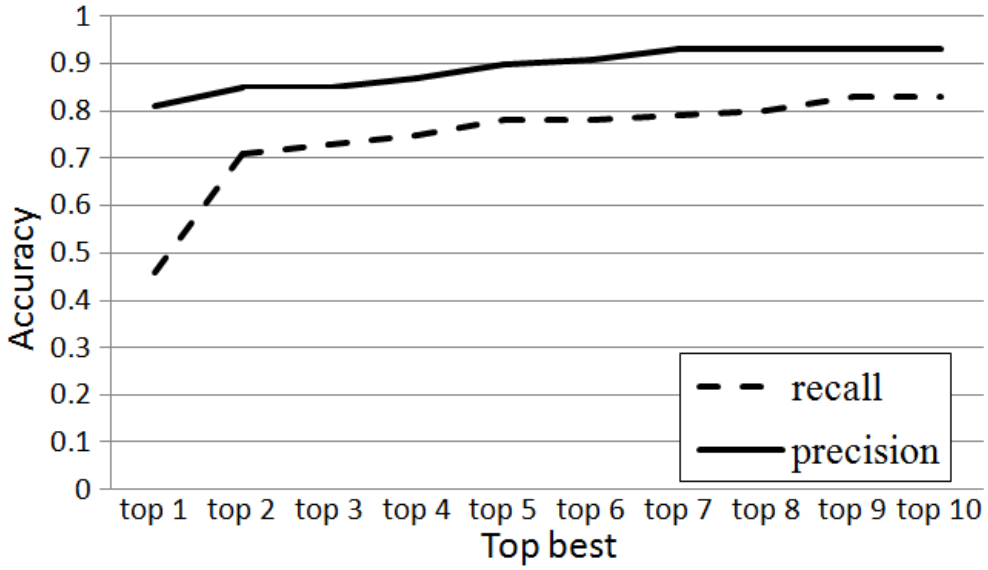


Рис. 2.10. Графік, що відображає результати точності та повноти з використанням метрики перекриття для розуміння відео в межах топ-10 результатів функціональних одиниць

Для кількісної оцінки впливу FOON обчислювався F-Score у двох сценаріях (рисунок 2.11):

- З FOON: F-Score обчислювався з використанням метрики перекриття для справжніх та ідентифікованих ФО.

- Без FOON: Метрика перекриття обчислювалася між найвище ранжованими об'єктами/класами руху та об'єктами/вузлами руху в справжній ФО.

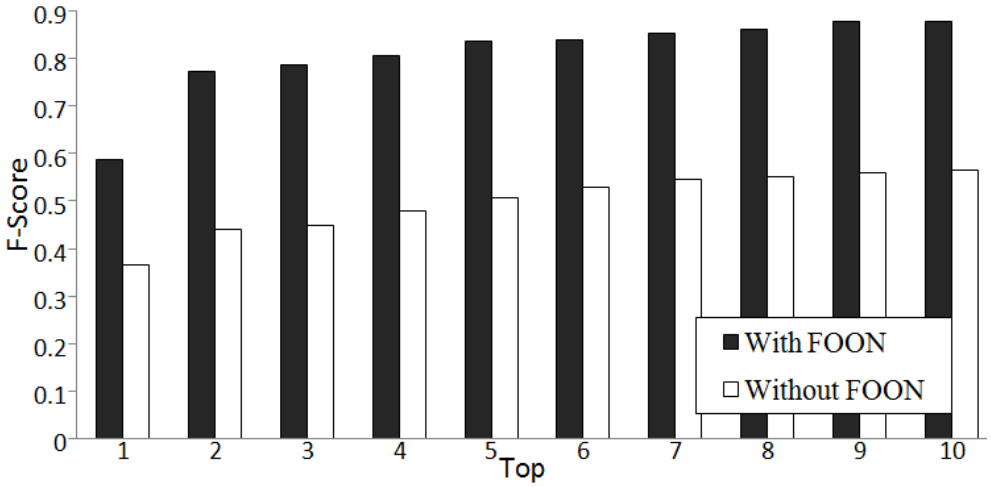


Рис. 2.11. Графік, що відображає обчислені F-Scores для розуміння відео з використанням FOON та без нього

Використання FOON забезпечило значно вищі F-Scores, оскільки вузли об'єктів та рухів у відео сприймаються краще, коли FOON використовується як структурована довідкова база.

2.6.2. Виведення завдання (класифікація рецептів)

Алгоритм застосовувався для класифікації рецептів для небачених кулінарних відео (тестовий набір з 8 відео). Рецепти в FOON були кластеризовані у 13 класів рецептів (наприклад, "Торт", "Омлет", "Салат").

Методологія наступна:

- Створення кластерів рецептів на основі всіх відео в навчальному наборі.

- Обчислення відстані подібності між тестовим відео та кожним кластером.

- Подібність відео з кластером агрегувалася з подібностей ФО та подібностей використаних об'єктів (порядок ФО не враховувався).

- Відео присвоювався клас рецепту з найвищою подібністю.

Для класифікації відео відповідно до рецепту, спочатку були створені кластери рецептів з використанням усіх відео в навчальному наборі. Обчислювалася відстань подібності (similarity distance) між поточним відео

та кожним кластером (рецептом), і обирався найближчий кластер як рецепт, асоційований із цим відео.

Щоб обчислити відстань подібності між поточним відео та кластером, обчислювалася середня подібність відео з кожним із відео в кластері. Подібність між відео та рецептом визначалася як агрегована подібність:

- подібність функціональних одиниць у відео з подібністю функціональних одиниць у рецепті.

- подібність використаних об'єктів у відео з подібністю об'єктних вузлів у рецепті.

Важливо, що при порівнянні подібності порядок функціональних одиниць не перевірявся. Клас рецепту з найвищою подібністю був присвоєний відео. Рисунок 2.12 демонструє ідентифіковані функціональні одиниці відео, що показує приготування локшини кухарем.

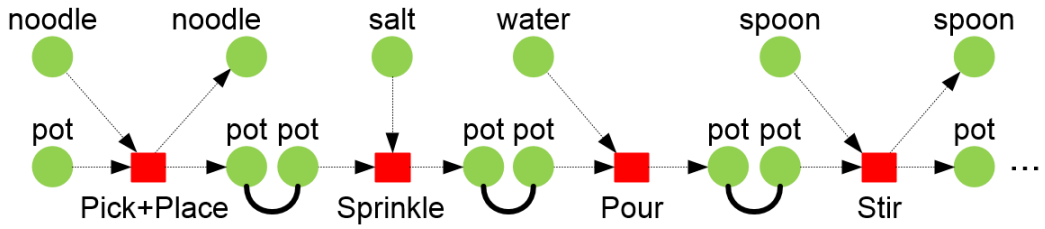


Рис. 2.12. Ілюстрація ідентифікованих функціональних одиниць для відео приготування локшини

Таблиця 2.2.

Результати класифікації рецептів

Використана процедура	Top 1	Top 2
Об'єкти, марковані вручну	37.5%	100%
Об'єкти, марковані автоматично	25%	75%

Таблиця 2.2 демонструє результати класифікації рецептів. Алгоритм, який використовує FOON, здатен наближено передбачити тип рецепту, що готується у відео, припускаючи коректну ідентифікацію об'єктів. Це також

підкреслює, що рух об'єктів може слугувати підказкою для типу активності рецепту.

На відміну від більшості робіт, які виводять активність єдиним реченням або міткою, наше дослідження виводить підграфи, що представляють атомарні активності. Це робить пряме кількісне порівняння з іншими методами складним. Проведений аналіз через метрику перекриття та порівняння конвеєрів з/без FOON підтверджує, що FOON є потужним представленням знань, здатним до розуміння відео.

Хоча поточна структура використовує інформацію з однієї камери в один момент часу, вона може бути інтегрована в багатовидові системи. Індивідуальні передбачення з різних камер можуть бути зібрані та об'єднані для отримання остаточного передбачення. Об'єднання також можливе на рівні впевненості, коли впевненості об'єктів з кожного виду агрегуються.

Основна мета — розуміння відео за допомогою представлення знань. Запропонований конвеєр є кроком до автоматичного розширення графа FOON.

Висновки до розділу

Другий розділ був присвячений формуванню методології інтелектуального аналізу кулінарного контенту на основі графових моделей знань і сучасних архітектур глибинного навчання. У роботі було детально обґрунтовано доцільність використання функціональної об’єктно-орієнтованої мережі як основи для структурування кулінарних процесів. Встановлено, що така мережа дозволяє інтегрувати інформацію про об’єкти, дії, стани та причиново-наслідкові зв’язки, забезпечуючи зрозуміле описання кулінарних задач. Було запропоновано архітектуру, що поєднує механізми виявлення функціональних об’єктів, рухів та функціональних одиниць у єдиний конвеєр розуміння відео. Показано, що використання Faster R-CNN для розпізнавання функціональних об’єктів забезпечує високу точність в

умовах різноманітності кулінарних сцен. Аналіз процесу виявлення функціональних одиниць продемонстрував, що ймовірнісне оцінювання підвищує стабільність та надійність моделі на складних даних. Було описано механізм побудови графа завдань, який дозволяє інтерпретувати відео кулінарних процесів не лише як набір дій, а як логічно пов’язану структуру.

РОЗДІЛ 3. РЕАЛІЗАЦІЯ МЕТОДІВ ТА ПРЕДСТАВЛЕННЯ МЕТОДОЛОГІЇ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КУЛІНАРНОГО КОНТЕНТУ

3.1. Моделювання станів об'єктів у кулінарному контексті

Ефективне робототехнічне планування завдань та маніпулювання вимагає глибокого розуміння як об'єктів, так і змін їхнього стану в маніпульованій сцені. У попередніх розділах було зосереджено увагу на аналізі об'єктів, рухів та сцен для розуміння відео в кулінарному контексті. Хоча багато рухів, як-от нарізання, неявно містять інформацію про зміну стану об'єкта, для явного використання моделювання стану в розумінні завдань маніпулювання необхідно експліцитно кодувати стан кулінарного об'єкта (наприклад, "нарізана цибуля").

Попри високу активність досліджень у сфері розуміння зображень, розпізнавання об'єктів та розуміння сцени, ідентифікація станів об'єктів досі не привернула достатньої уваги в галузях комп'ютерного зору та робототехніки. Стан об'єкта можна визначити як спостережувані характеристики, в які об'єкт може бути трансформований внаслідок діяльності суб'єкта (людини або робота). Стан може бути описаний через такі атрибути, як форма, текстура або колір.

Наприклад, помідор може існувати в багатьох станах: "цілий", "нарізаний", "подрібнений". У послідовності кулінарної активності цілий помідор може бути спочатку нарізаний, а потім подрібнений.

Для інтелектуального робота-кухаря необхідність планувати свої рухи залежно від початкового стану об'єкта є першочерговою. Якщо робот отримує цілий помідор для приготування салату, він повинен запланувати послідовність дій ("помити", "нарізати", "подрібнити"). Якщо ж він отримує вже нарізаний помідор, початкові кроки пропускаються, і він виконує лише "подрібнення". Таким чином, для детального розуміння людської активності,

планування завдань робота та контролю маніпулювання критично важливо не лише розпізнати об'єкт, але й ідентифікувати його поточний стан.

Різні стани об'єкта та переходи між ними вимагають специфічних маніпуляцій та типів захоплення. Наприклад, ціла морква захоплюється інакше, ніж терта або нарізана морква. Утримання цілої моркви для нарізання, половини моркви для тертя або моркви, нарізаної соломкою, для подрібнення — кожен із цих сценаріїв вимагає унікальних типів захоплення. Отримання зворотного зв'язку від навколишнього середовища в режимі реального часу щодо стану об'єкта є необхідним для вирішення типу захоплення.

У цьому розділі ми:

- Надаємо формальне визначення станів об'єктів та пропонуємо таксономію станів для систематичного аналізу кулінарних змін.

- На основі таксономії створюємо та представляємо громадськості набір даних кулінарних станів.

- Пропонуємо моделі для спільного розпізнавання об'єктів та їхніх станів.

- Аналізуємо набір даних Recipe1M на предмет кулінарних станів та створюємо моделі для кодування наборів станів на даному зображенні.

- Експериментально демонструємо, що використання кодування станів суттєво покращує розпізнавання набору інгредієнтів зі зображення.

3.2. Набір даних для спільного розпізнавання об'єктів та їхніх станів у кулінарному контенті

У цьому розділі представлено обґрунтування, методологію збору та детальний аналіз набору даних ідентифікації станів, розробленого для вирішення виклику в комп'ютерному зорі та робототехніці.

У повсякденній діяльності люди виконують завдання, враховуючи як об'єкти, так і їхні стани, а також динаміку їхньої взаємодії. Для ідеального

маніпулювання навколишнім середовищем інтелектуальна система повинна володіти точними знаннями про об'єкти, їхні функціональні можливості та поточний стан. Оскільки об'єкт може мати різноманітні форми та стани, це обумовлює застосування різних стратегій маніпулювання.

Приклад (нарізання перцю кубиками). Проста дія, як-от нарізання перцю кубиками, вимагає послідовного переходу через кілька станів (наприклад, цілий → половина → соломка → кубики). Робот повинен отримувати постійний зворотний зв'язок щодо поточного стану об'єкта, щоб адаптувати наступні кроки маніпулювання. Таким чином, знання поточного стану об'єкта є критично важливим для формування ефективного підходу до маніпуляції.

Ми визначаємо стан об'єкта (наприклад, помідора) як різні фізичні форми ("кубики", "паста", "сік" або "цілий"), в які об'єкт може бути трансформований внаслідок діяльності суб'єкта.

Вирішення виклику ідентифікації стану спрямоване на підвищення точності розуміння та виконання робототехнічних завдань маніпулювання, таких як захоплення. З огляду на важливість кулінарної сфери для робототехнічних систем, дане дослідження зосереджено на кулінарних об'єктах.

Аналіз станів, отриманий зі статистичних даних представлення знань, виявив дві основні категорії станів кулінарних об'єктів:

- зміна форми включає підстани розділений (separated), трансформований (transformed) та об'єднаний (combined).

- зміна поверхні: Включає підстани зміна кольору та зміна текстури.

На рисунку 3.1 зображено ієрархічне представлення досліджених станів, що містить загалом 22 деталізовані підстани. Ці стани охоплюють простір станів усіх важливих об'єктів.

Для даної проблеми було обрано лише 12 репрезентативних станів (показані сінім кольором на рисунку 3.1) через обмежену кількість навчальних зразків зображень для виключених станів. У цьому дослідженні

для спрощення припускається, що об'єкт має лише один стан у певний момент часу (на одному знімку).

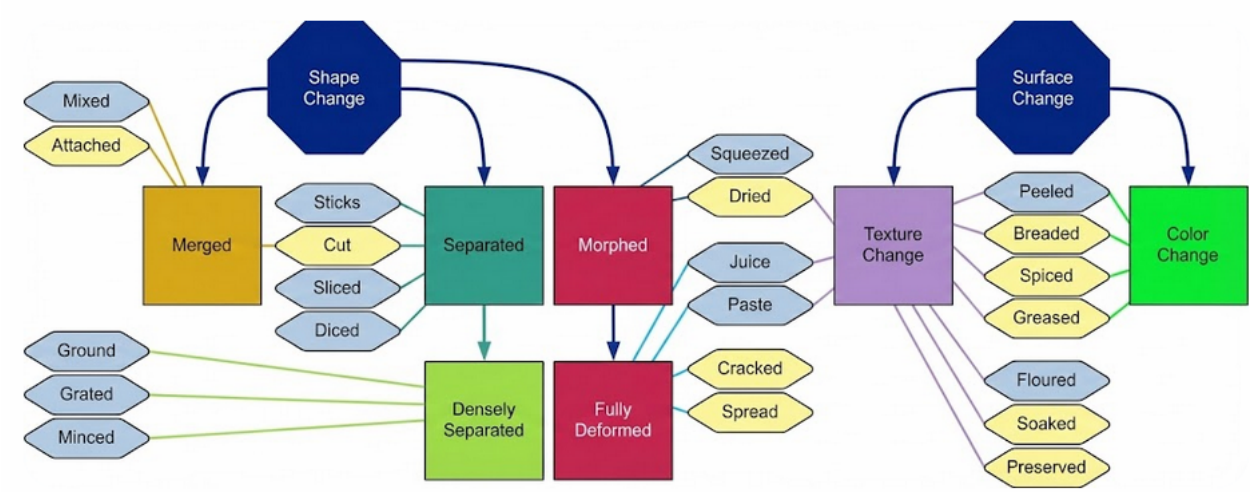


Рис. 3.1. Ілюстрація таксономії станів, створеної шляхом аналізу представлення знань для задачі ідентифікації станів кулінарного контенту

Зображення для набору даних збиралися через пошуковий механізм Google за допомогою комбінації ключових слів, що включали кожен об'єкт та його стан (наприклад, "помідор" та "нарізаний"). Зібрані зображення були відфільтровані від нерелевантного контенту (наприклад, мультфільмів). Анотація проводилася вручну за допомогою інструментів Vatic та labelbox, з подальшою перевіркою міток.

Vatic — це потужний програмний інструмент з відкритим вихідним кодом, спеціально розроблений для анотації (маркування) відео з метою навчання систем комп'ютерного зору. Він є одним із перших широко відомих інструментів, що використовували модель людських обчислень (human computation), застосовуючи принцип Amazon Mechanical Turk (MTurk) для масштабування процесу.

Зібраний набір даних включає:

- 17 основних кулінарних об'єктів (наприклад, помідор, цибуля, м'ясо, сир, тісто).

- 11 класів станів: цілий, очищений, в борошні, нарізаний, нарізаний кубиками, тертий, соломкою, сік, кремовий, змішаний, інший.

Загальний обсяг набору даних становить 9309 зображень: 6498 (70%) для навчання, 1413 (15%) для тестування та 1398 (15%) для валідації. Статистика класів (рисунок 3.2) показує, що класи "цілий" та "нарізаний" містять понад 1000 зображень, тоді як інші класи мають приблизно 700–1000 зображень.

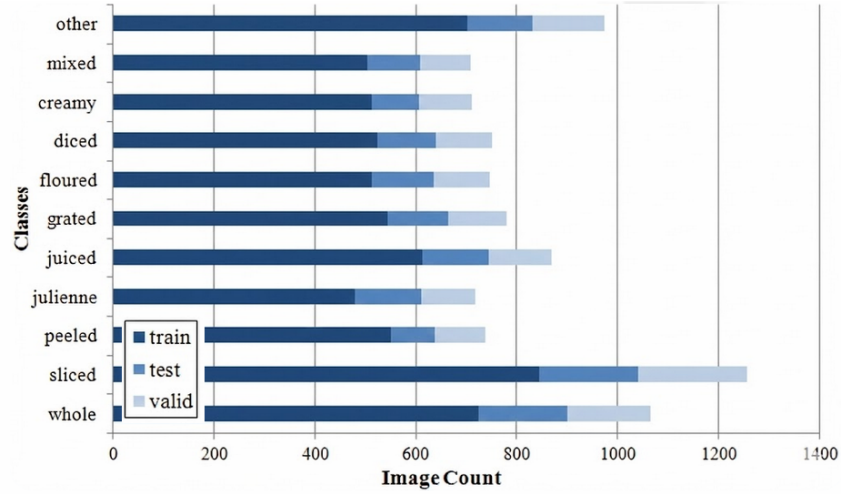


Рис. 3.2. Ілюстрація статистики одинадцяти класів станів у кулінарному наборі даних, що містить понад 10 000 зображень

Надано стислі визначення для 11 класів станів для забезпечення чіткої межі між категоріями (рисунок 3.3).

Рис. 3.3. Ілюстрація статистики одинадцяти класів станів у кулінарному наборі даних

Таблиця 3.1.

Опис набору даних

Стан	Визначення	Приклади
Цілий	Об'єкти в їхньому оригінальному форматі та формі.	Цілий перець, ціла курка.
Очищений	Об'єкти, з яких видалено зовнішню оболонку, але які не нарізані чи трансформовані.	Очищене яйце, цибуля, часник.
В Борошні	Об'єкти, вкриті борошном.	Об'єкти у борошні.
Тертий	Об'єкти, щільно розділені на дрібні фракції.	Панірувальні сухарі, подрібнений часник.
Соломка	Об'єкти, нарізані довгими та тонкими смужками/паличками.	Морквяні палички, картопля фрі.
Нарізаний Кубиками	Об'єкти, нарізані кубиками або подрібнені.	Нарізана кубиками цибуля/помідор, кубики масла/сиру.
Нарізаний	Об'єкти, тонко нарізані (на скибки).	Нарізана морква/перець/цибуля, скибки м'яса/сиру. *Об'єкти, нарізані іншим способом (навпіл, кубиками), виключаються.*
Сік	Рідкий стан об'єктів.	Молоко, розтоплене масло, томатний сік.
Кремовий	Об'єкти з кремовою текстурою.	Вершки, кремове масло, томатна паста, картопляне пюре.
Змішаний	Суміш кількох об'єктів.	Салати (як суміш інгредієнтів).
Інший	Будь-який стан, не охоплений попередніми категоріями.	Картопля, нарізана навпіл, вичавлений лимон, зображення з кількома станами.

3.3. Застосування глибокої архітектури ResNet для базового розпізнавання станів об'єктів в кулінарному контенті

У цьому розділі представлено аналіз запропонованого набору даних станів об'єктів з використанням базової моделі глибокого навчання, а також надано експериментальні результати.

3.3.1. Архітектура глибокої мережі та навчання

Для вирішення проблеми класифікації станів було обрано глибоку згорткову базову модель ResNet (Residual Network). Базова модель використовує архітектуру ResNet до 46-го шару активації. До цієї основи були додані наступні шари:

- Один шар згортки 1×1 (для зменшення глибини карт ознак).

- Два шари згортки (для захоплення нових просторових ознак, специфічних для ідентифікації станів).

- Шар глобального усереднення.

- Кінцевий шар Softmax з 11 вихідними класами.

У кожному шарі застосовувалася пакетна нормалізація (Batch Normalization) для нормалізації та регуляризації, а також додавалася випадковість (dropout) для запобігання перенавчанню.

Структура мережі подана на рисунку 3.4.

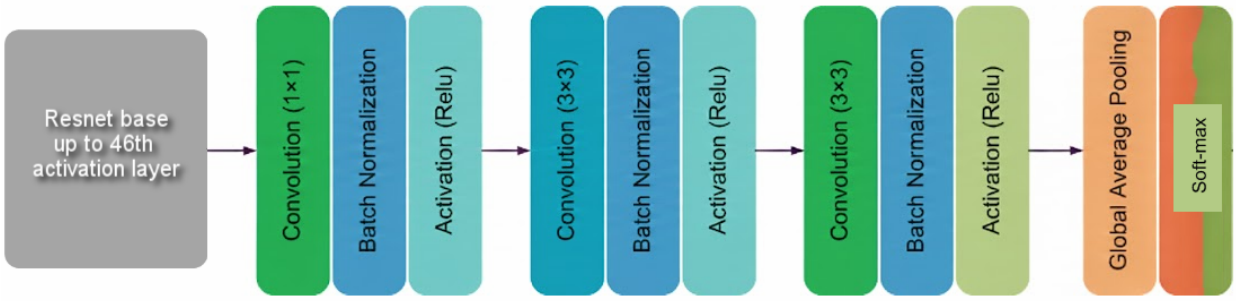


Рис. 3.4. Запропонована структура базової мережі

На рисунку 3.4 мережа включає основу ResNet, за якою слідують три шари згортки (convolution), пакетної нормалізації (batch-norm) та ReLU

Для ініціалізації ваг була використана модель ResNet, попередньо навчена на ImageNet:

Етап 1 (заморожування) - спочатку попередньо навчені ваги базової ResNet були заморожені, і навчалися лише додані шари.

Етап 2 (тонке налаштування), тобто на пізніших етапах навчання вся мережа (включаючи базу ResNet) була тонко налаштована.

Було розроблено три основні експерименти:

1. Навчання та тестування глибокої архітектури на всьому наборі даних C-SID.

2. Покращення моделі шляхом тонкого налаштування окремо для кожного об'єкта.

3. Тестування моделі на вибірці зображень ImageNet.

Модель навчалася за допомогою оптимізатора Adam (learning rate=0.001). Навчання включало 10 епох із замороженою базою ResNet, після чого слідувало 25 епох тонкого налаштування всіх шарів (learning rate=0.000005). Для зменшення перенавчання застосовувалися онлайн-розширення даних, L2-регуляризація та пакетна нормалізація.

Середня точність класу для навчального та тестового наборів становила 81.4% та 80.4% відповідно (таблиця 3.2).

Таблиця 3.2.

Базова точність класифікації для Top 2 результатів на наборі даних станів та підмножині ImageNet

	Модель	**Набір даних станів**	**Підмножина ImageNet**
		Top 1	Top 2
1	**Модель на основі ResNet** (Resnet-based Model)	80.4%	91.5%
2	Voting	82%	92%

За допомогою зваженого голосування між трьома навченими моделями точність розпізнавання станів зросла до 82%.

Також виконано тонке налаштування для об'єктів. Враховуючи сильну кореляцію між станом та типом об'єкта (наприклад, масло не може бути тертим), було розроблено 17 окремих моделей, тонко налаштованих для кожного об'єкта. Оскільки кожен об'єкт має різну кількість можливих станів (наприклад, часник має 5 станів, морква — 7), кінцевий шар Softmax був замінений відповідно до кількості станів, специфічних для об'єкта.

Для процедури тонкого налаштування використовувався 4-етапний процес навчання:

- Етапи 1–2: Заморожування всіх шарів, крім останнього.
- Етап 3: Розморожування додаткових шарів.
- Етап 4: Розморожування всієї моделі та повне тонке налаштування.

Точності класифікації для тонко налаштованих моделей, специфічних для кожного об'єкта (за винятком тіста), наведено у таблиці 3.3.

Таблиця 3.3.

Базова точність класифікації станів на основі тонкого налаштування індивідуального інгредієнта та кількість станів на інгредієнт

Об'єкт	**Top 1**	**Голосування (Voting)**	**Стани (States)**	**Тестовий набір (Test Set)**
mushroom (гриб)	94.6%	97.8%	3	42
onion (цибуля)	81.2%	85%	7	87
strawberry (полуниця)	91.6%	92%	4	68
bread (хліб)	77.9%	78.9%	6	120
butter (масло)	68.7%	72.7%	5	60
carrot (морква)	77.4%	83.9%	8	132
egg (яйце)	90.6%	89.2%	5	85
garlic (часник)	86.7%	85.3%	5	75
lemon (лимон)	90.7%	94.9%	6	109
milk (молоко)	100%	100%	2	40
pepper (перець)	96.1%	97.5%	5	76
potato (картопля)	84%	88.3%	8	106
tomato (помідор)	88.5%	92.1%	7	103
cheese (сир)	82.7%	78.7%	4	75
beef/pork (яловичина/свинина)	86.7%	86.7%	5	60
chicken (курка)	88.8%	88.4%	6	109
average (середнє)	86%	88%	**5**	**86**

Для додаткової валідації та внеску в набір даних ImageNet було проведено тестування моделі на підмножині ImageNet. Було відібрано по 50 зображень із 16 категорій об'єктів ImageNet (за винятком яловичини/курки), що становило 500 зображень, які були вручну позначені 11 класами станів.

Середня точність ідентифікації стану на підмножині ImageNet склала 78 %. Індивідуальні точності для топ-1, 2 та 3 результатів наведено у таблиці 3.4.

Точність класифікації всіх класів на тестовому наборі (за винятком класу "інший") становила мінімум 70%. Клас "інший" продемонстрував найнижчу точність через високу різноманітність вмісту (наприклад, страви, комбінації різних станів).

Більшість помилок моделі були спричинені неоднозначними або багатозначними зображеннями. Це підкреслює, що розгляд зображення як цілого є обмеженим, і що виявлення всіх станів всередині зображення, а не лише його класифікація, є необхідним напрямком для покращення.

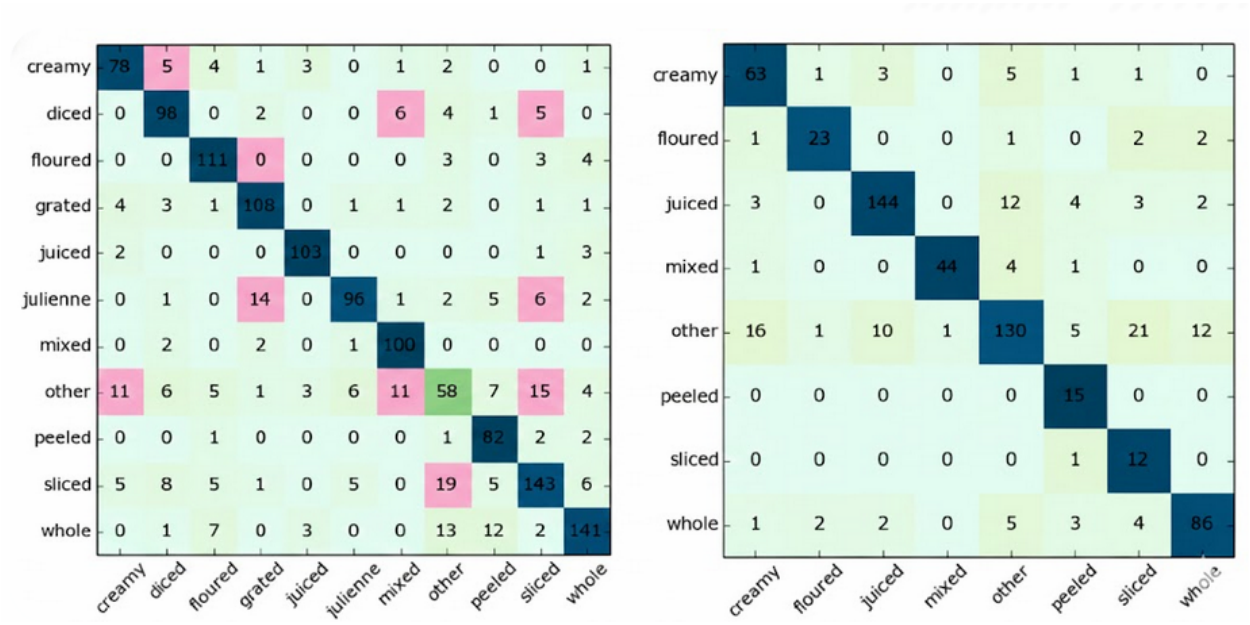
Незважаючи на наявність більшої кількості неоднозначних зображень у підмножині ImageNet, середня точність (78%) була лише трохи нижчою, ніж на нашому власному наборі даних.

Аналіз помилок показав, що модель здатна вилучати релевантні ознаки (наприклад, передбачати очищений для напівочищеного перцю), але не має інструменту для ідентифікації кількох станів у зображенні одночасно. Наприклад, зображення, що містить кілька станів, було зараховане як неправильне передбачення, хоча правильні стани були у топ-3 передбачень моделі. Матриця плутанини для ImageNet (рисунок 3.5 б) також показала відсутність деяких класів через обмежену кількість зразків.

3.3.2. Аналіз помилок на основі матриць плутанини

Це підтверджує, що для досягнення оптимальних результатів необхідний перехід до архітектури, здатної виконувати спільне виявлення та розпізнавання станів. Матриці плутанини представлені на рис. 3.5.

використовуються для оцінки продуктивності базової моделі класифікації станів на двох різних наборах даних



а) на множині даний ідентифікації станів б) на підмножині ImageNet

Рис. 3.5. Матриці плутанини (невідповідності)

Матриця на рис. 3.5 а ілюструє продуктивність моделі, навченої та протестованої на оригінальному наборі даних, зібраному для цієї задачі.

Розмірність: 11×11, що відповідає 11 класам станів (наприклад, creamy, diced, whole, other).

Основні діагональні елементи (висока точність): Темно-сині значення по головній діагоналі (наприклад, 78 для creamy, 98 для diced, 141 для whole) вказують на кількість екземплярів, які були правильно класифіковані. Ці високі значення підтверджують високу загальну точність моделі (80 %, згідно з таблицею 3.2).

Клас "other" демонструє найвищу помилку (розсіяння). Він має найбільшу кількість неправильних передбачень в інших класах (рядок other) і водночас найбільшу кількість помилкових передбачень з інших класів (стовпець other). Наприклад, 19 зразків, які насправді є sliced (нарізаний), були помилково класифіковані як other. Між класами "diced" (нарізаний

кубиками) та "julienne" (соломка) існує невелика плутанина, що свідчить про їхню схожість.

Отже, модель демонструє високу точність для більшості класів, що добре розділяються, але має значні труднощі з класифікацією неоднорідного та багатозначного класу "other".

Матриця на рис. 3.5 б (ImageNet) показує, наскільки добре та сама модель узагальнює свої знання на зовнішньому, менш спеціалізованому наборі даних. Ця матриця має меншу розмірність (8×8), оскільки вона не включає деякі класи (такі як julienne, diced, grated), через брак репрезентативних зразків цих станів у підмножині ImageNet.

Діагональні елементи все ще досить високі (наприклад, 63 для creamy, 86 для whole), підтверджуючи загальну точність близько 78%. Це вказує на хорошу здатність моделі до перенесення знань.

На ImageNet посилюються проблеми з неоднорідними класами:

- Клас "other" продовжує демонструвати значну плутанину.

- Спостерігається помітна плутанина між класами, що вказує на більшу неоднозначність зображень ImageNet порівняно з цільовим набором даних. Наприклад, деякі зразки, які є peeled (очищений), помилково класифіковані як other або sliced.

Як висновок, можна вказати, що продуктивність моделі залишається високою, але трохи знижується, що є очікуваним при тестуванні на більш неконтрольованому та різноманітному наборі даних. Нестача деяких класів станів у ImageNet обмежує повноту оцінки.

3.4. Методологія передбачення інгредієнтів багатоскладових страв на основі їх зображень

Традиційні підходи до класифікації зображень часто припускають наявність лише одного об'єкта або стану, що не відповідає реаліям багатокомпонентних страв. У цьому розділі розглядається більш реалістична

постановка задачі: ідентифікація кількох інгредієнтів та їхніх станів на одному зображенні готової страви. Ми досліджуємо, як візуальне розрізнення станів може покращити продуктивність моделі передбачення інгредієнтів.

В експериментах використовується набір даних Recipe1M, який не містить явних міток станів. Ми описуємо методологію виведення станів у Recipe1M, процес створення їхніх вбудовувань та архітектуру моделі для передбачення інгредієнтів на основі інтегрованих ознак стану.

Рис. 3.6. Зображення морквяно-імбирного супу та асоційовані з ним інгредієнти (наприклад, цибуля) і їхні стани (наприклад, нарізаний)

Стани та токени (візуально значущі прикметники/дієслова) були виведені з текстових описів Recipe1M за трьома етапами:

- Основний набір (base set): визначення початкового набору класів станів (наприклад, змішаний, тертий, очищений).

- Вилучення псевдонімів (alias extraction): ручне та напівавтоматичне вилучення синонімів (псевдонімів) станів.

- Додавання вручну (manual addition): аналіз дієслів у минулому часі та високочастотних прикметників з описів інгредієнтів. Якщо вони візуально значущі та схожі на основні стани — додаються як псевдоніми; якщо ні, але мають високу частоту — додається новий клас стану.

Щоб уникнути упередження, іншим набором токенів стали всі прикметники та дієслова з одним домінантним візуальним значенням (після

видалення слів без візуального ефекту, наприклад, солений). Менш часті слова були об'єднані в клас "невідомо".

Для кодування візуальної інформації про стани та токени в числові вектори (вбудовування) застосовується модель, на основі декодера трансформера.

1. Вилучення ознак зображення. Зображення I кодується за допомогою кодувальника ResNet у карти ознак $G \in RC \times S \times S$, які потім сплощуються до $F \in RK \times D$.

2. Генерація векторів імовірності. F подається на вхід декодеру трансформера, який покроково передбачає вектори логітів стану st. Для моделювання залежностей, передбачені стани знову подаються на вхід.

3. Усунення порядку та пудлінг. Передбачені стани об'єднуються в матрицю S. До S застосовується операція max-pooling (усуваючи позиційне кодування), що створює вектор імовірностей $Ps \in RN \times 1$ (де N — кількість класів станів).

4. Фінальне вбудовування. Ps кодується через повноз'єднаний шар у вектор вбудовування стану $Es \in RK \times 1$.

Такий самий процес застосовується для отримання вбудовування слів (токенів) $Ew \in RK \times 1$. Ці вектори Es та Ew містять візуальну асоціацію вхідного зображення з кожним із можливих станів/токенів.

Мета передбачення інгредієнтів на основі станів це використання вбудовувань станів Es та токенів Ew для підвищення точності передбачення списку інгредієнтів за зображенням готової страви I.

На першому етапі вбудовування станів та токенів об'єднуються з візуальними ознаками зображення F (сплощені карти ознак ResNet):

$$Z = [F, E_s, E_w]$$

де $Z \in R^{K \times (D+Ne)}$ — об'єднані ознаки (N_e=2).

2. Матриця Z подається на вхід декодеру трансформера. Цей декодер використовує механізм багатопотокової уваги скалярного добутку (multi-head self-attention):

$$Z_{att} = \text{softmax}\left(\frac{ZQ^T}{\sqrt{d_q}}\right)Z$$

де Q — матриця запитів, що складається з попередньо передбачених вбудовувань інгредієнтів.

3. Інгредієнти передбачаються покроково через декодер трансформера, причому попередні передбачення подаються як вхідні вбудовування.

4. Фінальна класифікація. Після проходження шарів уваги застосовується шар max-pooling у часі для усунення порядку. Це генерує фінальний вектор імовірностей, на якому навчається функція втрат бінарної перехресної ентропії (Binary Cross-Entropy) для моделювання інгредієнтів без урахування порядку.

Використання вбудовувань станів замість one-hot кодування дозволяє уникнути ускладнення моделі та потреби навчати велику кількість параметрів вбудовувань окремо. Абстрактна ілюстрація моделі наведена на рисунку 3.7.

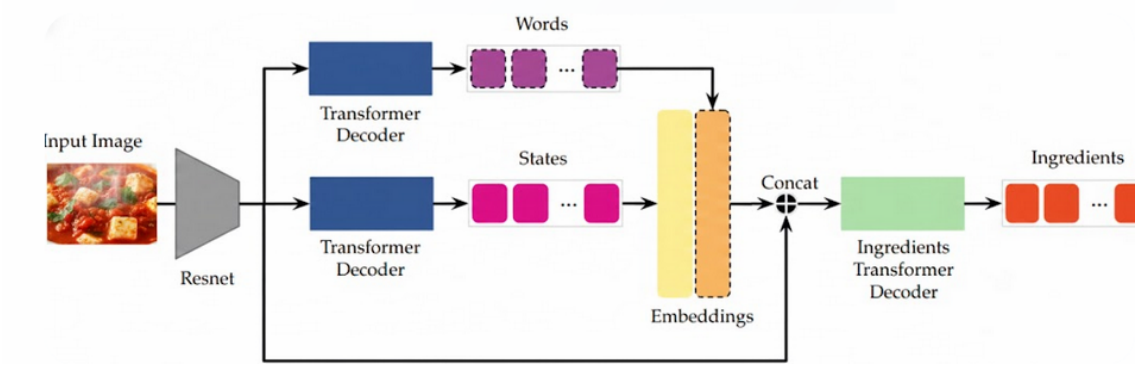


Рис. 3.7. Двоетапна модель, яка передбачає інгредієнти за одним зображенням

На першому етапі, маючи зображення готової страви, створюються два набори вбудовувань: станів та слів. На другому етапі, використовуючи

вбудовування станів та вбудовування зображення, інгредієнти передбачаються покроково за допомогою декодера трансформера.

Для навчання як моделей вбудовування, так і моделі передбачення інгредієнтів, використовується комбінація трьох функцій втрат:

- Функція втрат бінарної перехресної ентропії.
- Функція втрат кінцевого токена.
- Функція втрат кардинальності (кількості передбачень).

Отже, було запропоновано низку алгоритмів для екстракції знань, виведення та розуміння візуального кулінарного контенту. Під візуальним кулінарним контентом ми розуміємо зображення та відео, що фіксують приготування їжі або готові страви. Для ефективної екстракції знань із відео необхідно застосовувати сучасні методи комп'ютерного зору, зокрема, алгоритми глибокого навчання.

Ми розробили конвеєр, що поєднує глибокі згорткові мережі (DCN) та авторегресійні рекурентні нейронні мережі (RNN) для досягнення комплексного розуміння відео, яке зображує кулінарне завдання.

Висновки до розділу

У третьому розділі було реалізовано методи та інструменти, що забезпечують практичне застосування розробленої методології інтелектуального аналізу кулінарного контенту. Першою вирішеною задачею було моделювання станів кулінарних об’єктів, що дозволило створити базу для подальшого навчання моделей глибинного аналізу. Було сформовано та проаналізовано спеціалізований набір даних, орієнтований на спільне розпізнавання об’єктів та їхніх станів, що значно підвищує можливості моделі в реальних сценаріях. Застосування архітектури ResNet для класифікації станів об’єктів продемонструвало достатньо високу точність за умови правильної підготовки даних і збалансованості класів. Проведений аналіз матриць плутанини виявив найбільш проблемні категорії станів, що

відкриває можливості для подальшої оптимізації мережі. Було показано, що помилки здебільшого пов’язані зі схожими візуальними характеристиками об’єктів, що є типовою проблемою в кулінарних даних. Розроблена методика передбачення інгредієнтів на основі зображення багатоскладових страв довела ефективність використання глибинних моделей для високорівневої семантичної інтерпретації. Здійснена реалізація підтвердила, що поєднання візуальних методів із графовими представленнями знань дозволяє отримати комплексний підхід до інтелектуального аналізу кулінарного контенту. Було продемонстровано, що модель здатна узагальнювати знання та застосовувати їх до нових прикладів, що свідчить про її практичну придатність.

ВИСНОВКИ

У магістерській роботі було проведено дослідження методів, моделей та інструментів інтелектуального аналізу кулінарного контенту, що охоплює зображення, відео та структуровані знання про об’єкти, дії та стани в кулінарних процесах. Робота узагальнює сучасні тенденції розвитку технологій комп’ютерного зору, глибинного навчання та графових представлень знань, орієнтованих на автоматизацію кулінарних завдань, інтерпретацію рецептів, оцінювання інгредієнтів та розуміння відеопроцесів приготування їжі.

Проведений аналіз предметної області дозволив встановити, що кулінарний контент є складним інформаційним середовищем, яке поєднує візуальні, семантичні та процедурні складові. Для його повноцінного опрацювання необхідні моделі, здатні інтегрувати інформацію про об’єкти, їхні властивості, функціональні ролі, часові залежності, а також процедури перетворень, характерні для кулінарних сценаріїв. Огляд наукових робіт показав, що класичні підходи комп’ютерного зору не забезпечують достатньої інтерпретовності та структурності знань, тоді як графові моделі, функціонально-об’єктні мережі та багаторівневі глибинні архітектури здатні суттєво підвищити точність і семантичну насиченість обробки.

На основі аналізу сучасних методологій було запропоновано та обґрунтовано застосування функціональної об’єктно-орієнтованої мережі як апріорного представлення знань про кулінарні дії та об’єкти. Такий підхід дозволяє об’єднати інформацію про інгредієнти, кухонний інвентар, стани об’єктів та функціональні дії у єдину графову структуру, що суттєво підвищує точність інтерпретації візуального контенту. Розроблений чотириетапний конвеєр розуміння кулінарного відео забезпечує послідовне перетворення даних від виявлення об’єктів до ідентифікації функціональних одиниць та побудови графа задач, що дозволяє відтворювати логіку рецептурних процесів на основі відео.

Для реалізації запропонованого підходу було проведено моделювання станів кулінарних об’єктів, сформовано спеціалізований набір даних та розроблено базову архітектуру на основі ResNet для розпізнавання станів інгредієнтів. Отримані результати підтвердили ефективність глибинних моделей у задачах аналізу кулінарного контенту, а здійснений аналіз помилок дозволив визначити найбільш проблемні класи станів та фактори, що впливають на якість класифікації. Додатково була запропонована методика прогнозування інгредієнтів багатоскладових страв за їхнім зображенням, що свідчить про практичний потенціал систем інтелектуального аналізу у сфері харчових технологій.

Комплексна оцінка конвеєра розуміння кулінарного відео засвідчила, що інтеграція графових представлень знань із сучасними алгоритмами комп’ютерного зору забезпечує значне підвищення точності та семантичної глибини аналізу порівняно з класичними нейромережевими підходами. Запропонована методологія може бути використана як основа для створення інтелектуальних кухонних асистентів, систем автоматичного генерування рецептів, рекомендаційних систем, а також робототехнічних комплексів, орієнтованих на автоматизацію побутових або промислових кулінарних процесів.

Отже, результати дослідження доводять доцільність використання структурованих знань, графових моделей та багаторівневих глибинних архітектур для інтелектуального аналізу кулінарного контенту. Сформована методологія забезпечує багатокомпонентне розуміння зображень і відео, а також відкриває можливості для подальших наукових і прикладних розробок у галузі автоматизації кулінарних процесів та обробки мультимодального контенту. Робота створює теоретичне та практичне підґрунтя для інноваційних досліджень у межах комп’ютерного зору, видобування знань та інтелектуальних систем підтримки кулінарних рішень.

ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. NEIL: Extracting Visual Knowledge from Web Data / Xinlei Chen // https://xinleic.xyz/papers/iccv13.pdf
2. Understanding VGG: The Backbone of Image Recognition. – https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/
3. Multi-scale object detection in remote sensing imagery with convolutional neural networks. - https://www.researchgate.net/publication/324903264_ Multi-scale_object_detection_in_remote_sensing_imagery_with_ convolutio nal_neural_networks
4. Towards Real-Time Smile Detection Based on Faster Region Convolutional Neural Network. - https://www.researchgate.net/publication/324549019_ Towards_Real-Time_Smile_Detection_Based_on_Faster_Region_Convo- lutional_Neural_Network
5. RetinaNet: Fast and Accurate Object Detection Model. - https://viso.ai/deep-learning/retinanet/
6. Companion: Easy Navigation App for Visually Impaired Persons. - https://www.researchgate.net/publication/353697284_Companion_Easy_Na vigation_App_for_Visually_Impaired_Persons
7. The Single-Shot MultiBox Detector (SSD) model consisting of a VGG and... | Download Scientific Diagram - https://www.researchgate.net/figure/The-Single-Shot-MultiBox-Detector-SSD-model-consisting-of-a-VGG-and-additional_fig1_348769625
8. Functional Object-Oriented Network: Construction & Expansion / David Paulius // https://www.kavrakilab.org/2017-rss-workshop/paulius.pdf
9. Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. “COMET: Commonsense Transformers for Automatic Knowledge Graph Construction.” Proceedings of ACL, Florence, Italy, ACL Press, 2019, pp. 4762–4779.

10. Chen, J., Wu, H., Jiang, Y.-G., & Xue, X. “Learning to Recognize Cooking Activities Using Deep and Multimodal Features.” IEEE Transactions on Multimedia, IEEE, 2018, pp. 1–12.
11. Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., & Torresani, L. “Learning Cross-modal Embeddings for Cooking Recipes and Food Images.” CVPR, Honolulu, USA, IEEE, 2017, pp. 3020–3028.
12. Min, W., Liu, L., Li, Z., Jiang, S., & Jain, R. “A Survey on Food Computing.” ACM Computing Surveys, ACM, 2020, pp. 1–46.
13. Chen, J., Ngo, C.-W., & Jiang, Y.-G. “Deep Recipe Retrieval: Semantics-Aware Embedding of Cooking Recipes and Food Images.” ACM Multimedia, Seoul, South Korea, ACM, 2018, pp. 1020–1028.
14. Yaguchi, H., Kato, Y., & Yamasaki, T. “Mapping Cooking Actions in Video Using Hierarchical Neural Networks.” ICCV Workshops, Seoul, South Korea, IEEE, 2019, pp. 1–10.
15. Papadopoulos, D. P., Ferrari, V., & Schindler, K. “Video Understanding Through Structured Representations.” IJCV, Springer, 2021, pp. 1–24.
16. Jermsurawong, J., & Habash, N. “Predicting the Structure of Cooking Recipes.” EMNLP, Lisbon, Portugal, ACL Press, 2015, pp. 781–786.
17. Wu, X., Sahoo, D., & Hoi, S. C. “Recent Advances in Deep Learning for Object Detection.” Neurocomputing, Elsevier, 2020, pp. 31–60.
18. Misra, D., Bennett, A., Blukis, V., Niklasson, E., & Artzi, Y. “Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction.” EMNLP, Brussels, Belgium, ACL, 2018, pp. 2667–2680.
19. Zhao, S., & Wang, J. “Knowledge Graphs and Their Application in AI Systems.” IEEE Access, IEEE, 2019, pp. 176–190.
20. Dosovitskiy, A., et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” ICLR, Vienna, Austria, OpenReview, 2021.
21. Ren, S., He, K., Girshick, R., & Sun, J. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” NeurIPS, Montreal, Canada, 2015, pp. 91–99.

22. Simonyan, K., & Zisserman, A. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” ICLR, San Diego, USA, ICLR Press, 2015.
23. He, K., Zhang, X., Ren, S., & Sun, J. “Deep Residual Learning for Image Recognition.” CVPR, Las Vegas, USA, IEEE, 2016, pp. 770–778.
24. Gao, L., Xu, X., & Song, J. “Hierarchical Attention Networks for Video Captioning.” AAAI, New Orleans, USA, AAAI Press, 2018, pp. 6832–6839.
25. Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. “VideoBERT: A Joint Model for Video and Language Representation Learning.” ICCV, Seoul, South Korea, IEEE, 2019, pp. 7463–7472.
26. Nagrani, A., Arnab, A., & Schmid, C. “Attention Bottlenecks for Multimodal Fusion.” NeurIPS, Vancouver, Canada, 2021, pp. 1–12.
27. Zhou, L., Zhang, C., & Liu, C. “Transformers in Vision: A Survey.” ACM Computing Surveys, ACM, 2022, pp. 1–38.
28. Ahn, S., Khandelwal, U., & Riedl, M. “Guiding Neural Story Generation Using Knowledge Graphs.” AAAI, Honolulu, USA, AAAI Press, 2019, pp. 5981–5988.
29. Lin, T. et al. “Microsoft COCO: Common Objects in Context.” ECCV, Zurich, Switzerland, Springer, 2014, pp. 740–755.
30. Zhou, B., et al. “Temporal Segment Networks for Action Recognition.” ECCV, Amsterdam, Netherlands, Springer, 2016, pp. 20–36.
31. Shi, X., et al. “Convolutional LSTM Network for Video Prediction.” NeurIPS, Montreal, Canada, 2015, pp. 802–810.
32. Fang, H., et al. “From Captions to Visual Concepts and Back.” CVPR, Boston, USA, IEEE, 2015, pp. 1473–1482.
33. Onoro-Rubio, D., & López-Sastre, R. “Towards Perspective-Free Object Counting.” CVPR, Las Vegas, USA, IEEE, 2016, pp. 789–798.
34. Kawano, Y., & Yanai, K. “Food Image Recognition with Deep Convolutional Features.” ACM Multimedia, Orlando, USA, ACM, 2014, pp. 1115–1118.

35. Singh, M., & Lee, Y. J. “End-to-End Localization and Ranking for Relative Attributes.” ECCV, Zurich, Switzerland, Springer, 2014, pp. 753–768.
36. Li, X., Chen, Y., Hu, Z., & Yang, H. “Object Detection in Complex Kitchen Environments.” IEEE Access, IEEE, 2021, pp. 1123–1137.
37. Ma, M., Ren, Y., & Li, S. “Graph-Based Reasoning for Visual Understanding.” IJCV, Springer, 2021, pp. 632–651.
38. Kato, Y., & Yamasaki, T. “Action Recognition in Cooking Videos Using CNN-LSTM Architectures.” ICIP, Athens, Greece, IEEE, 2018, pp. 1–5.
39. Chen, S., et al. “Cross-Modal Recipe Understanding via Joint Embedding.” Pattern Recognition, Elsevier, 2020, pp. 107–118.
40. Chen, C., & Luo, J. “Recipe2Video: Generating Videos from Cooking Recipes.” ECCV Workshops, Munich, Germany, Springer, 2018, pp. 1–16.
41. Aizawa, K., et al. “Understanding Food Images.” IEEE Multimedia, IEEE, 2014, pp. 17–28.
42. Meyers, A., et al. “Im2Calories: Towards an Automated Mobile Vision Food Diary.” ICCV, Santiago, Chile, IEEE, 2015, pp. 1233–1241.