Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Програмної інженерії

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

Дослідження методів генерації підписів до зображень

Виконав:
студент 2 курсу групи ІПЗм-20-4

Закіров М. С.
(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного забезпечення

Тип програми Освітньо-наукова

Керівник доц. Бабій А. С.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. Кафедри З.В. Дудар

2022 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерні науки

Кафедра Програмна Інженерія

Рівень вищої освіти другий (магістерський)

Спеціальність 121 — Інженерія програмного забезпечення

(код і повна назва)

Тип програми освітньо-наукова програма

Освітня програма Інженерія програмного забезпечення

(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

«___» _____ 202_ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Закірову Максиму Сергійовичу

(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження методів генерації підписів до зображень»

затверджена наказом по університету від «__» _____ 202_ р. №____

2. Термін подання студентом роботи до екзаменаційної комісії «__» ____ 202_ р.

3. Вихідні дані до роботи: середовище проектування PyCharm, мова розробки Python, фреймворк PyTorch, набір даних MS COCO, бібліотеки для оцінювання результатів підписів до зображень cider і coco-caption.

4. Перелік питань, що потрібно опрацювати в роботі реферат, вступ, перелік скорочень, аналіз предметної галузі, метрики оцінки ефективності Image Captioning алгоритмів, вибір та опис роботи алгоритму, вибір та анліз набору даних для обраної моделі, дослідження отриманих результатів.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз предметної галузі	26.02.2022	виконано
2	Формування вимог до системи	14.03.2022	виконано
3	Проектування системи	21.03.2022	виконано
4	Розробка програмного забезпечення	24.03.2022	виконано
5	Тестування системи	15.04.2022	виконано
6	Впровадження системи	16.04.2022	виконано
7	Підготовка пояснювальної записки	20.04.2022	виконано
8	Підготовка презентації та доповіді	01.05.2022	виконано
9	Перевірка на плагіат	16.05.2022	виконано
10	Нормоконтроль	20.05.2022	виконано
11	Рецензування	22.05.2022	виконано
12	Занесення диплома в електронний архів	23.05.2022	виконано
13	Попередній захист	21.05.2022	виконано
14	Допуск до захисту у зав. кафедри	23.05.2022	виконано

Дата видачі завдання _____202_ р.

Студент _____

(підпис)

Керівник роботи _____ _____доц. Бабій А. С._____

(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка: 70 с., 2 табл., 10 рис., 22 джерел.

IMAGE CAPTIONING, PYTHON, PYCHARM, PYTORCH, MS COCO, RECURRENT NEURAL NETWORKS, CONVOLUTIONAL NEURAL NETWORK, FAST R-CNN, OBJECT RELATIONAL TRANSFORMER.

Об’єктом дослідження є методи генерації підписів до зображень.

Метою роботи є реалізація одного з алгоритмів, що дозволяє аналізувати об'єкти у зображені і створювати підпис для їх опису та порівняння отриманих метрик за іншими існуючими моделями.

У результаті роботи виконано аналіз існуючих алгоритмів і обрано та описано алгоритм, який дозволяє вирішати задачу з State-Of-The-Art продуктивністю на наборі даних Microsoft COCO.

Explanatory note consists of: 70 p., 2 tables, 10 pictures, 22 sources.

IMAGE CAPTIONING, PYTHON, PYCHARM, PYTORCH, MS COCO, RECURRENT NEURAL NETWORKS, CONVOLUTIONAL NEURAL NETWORK, FAST R-CNN, OBJECT RELATIONAL TRANSFORMER.

The object of research is the methods of generating captions to images.

The aim of the work is to implement one of the algorithms that allows you to analyze objects in the image and create a caption to describe them and compare the resulting metrics with other existing models.

As a result, the analysis of existing algorithms was performed and an algorithm was selected and described that allows to solve the problem of State-Of-The-Art performance on the Microsoft COCO data set.

Умови публікації пояснювальної записки

Я, Закіров Максим Сергійович

(прізвище, ім’я, по батькові)

студент(ка) групи ІПЗм-20-4 здобувач вищої освіти на другому (магістерському) рівні

кафедра програмної інженерії,

(повна назва кафедри)

заявляю: моя кваліфікаційна робота на тему
«Дослідження методів генерації підписів до зображень»,

(назва роботи)

що буде представлена до ЕК для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений (а) з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

CNN – Згорткова нейронна мережа

LTSM – Мережі довгої короткострокової пам'яті

RNN – Рекурентна нейронна мережа

RPN – Мережа регіональних пропозицій

ReLU – Зрізаний лінійний вузол

ВСТУП

Завдяки безперервному накопиченню інформації, люди можуть поступово пізнавати оточуючий їх світ. На сьогодні отримання, обробка і аналіз візуальної інформації є одним з багатьох засобів пізнання світу. Алгоритми Image Captioning (або алгоритми генерації підписів до зображень) – це напрям досліджень, алгоритмів і методів, який перетворює образи та картинки у знання та інформацію для розуміння людиною, і надає більш широкі можливості для аналізу і обробки вилученої інформації.

Підписи до зображень(або Image Captioning) — завдання забезпечення опису вмісту зображення природною мовою — перебувають на перетині комп'ютерного зору та обробки природною мовою [1]. Можливість автоматично описати зміст зображення наразі є дуже непростою задачею. Ще складніше її вона стає через потребу використовувати правильно, з морфологічної точки зору, сформульовані речення та інші конструкції, які неповинні обмежуватися конкретною мовою. Це дуже важливий виклик для алгоритмів машинного навчання, оскільки він зводиться до імітації чудової людської здатності стискати величезну кількість помітної візуальної інформації в описову мову. Таке завдання є суттєво важче ніж наприклад вже добре вивчена задача класифікації зображень або задача розпізнання об’єктів, на яких буле зосереджено головна увага спільноти комп’ютерного зору. Це обумовлюється тим, що опис повинен охоплювати не тільки об'єкти, що містяться в зображення, але й виражати, як ці об'єкти пов'язані один до одного, а також їхні атрибути і діяльність до яких вони причетні.

Автоматичне створення підписів до зображення є завданням, дуже близьким до серця розуміння сцени — однією з основних цілей комп’ютерного зору. Моделі генерації підписів не тільки повинні бути достатньо потужними, щоб вирішувати проблеми комп’ютерного зору, пов’язані з визначенням об’єктів на зображенні, але вони також мають бути здатними фіксувати та виражати їхні стосунки природною мовою. Що стосується комп’ютерного зору, то покращені

архітектури згорткової нейронної мережі та виявлення об’єктів сприяли покращенню систем підписів зображень. А що стосується обробки природної мови, то більш складні послідовні моделі, такі як рекурентні нейронні мережі, що базуються на увазі, так само привели до більш точного створення підписів.

Більшість систем підписів зображень використовують структуру кодера-декодера, де вхідне зображення кодується у проміжне представлення інформації в зображенні, а потім декодується в описову текстову послідовність. Але також існують і інші моделі, які можуть бути актуальними для вузького колу задач, і які згадуються в даній роботі.

Задача Image Captioning і питання які вона вивчає, є дуже актуальними не тільки тому, що вони напряму допомагають людям з обмеженими можливостями пізнавати оточуючий світ, але й тому що вони має потенційну користь для більш якісного пошуку інформації у мережах, що наразі, у часи перенасичення інформацією звідусіль, є однією з найактуальніших проблем сучасної людини. З розвитком цього напряму, з’являються Підписи зображень вже використовуються у різноманітних програми, таких як: рекомендації в додатках для редагування, використання у віртуальних помічниках, для індексації зображень, для людей із вадами зору, для соціальних мереж та кілька інших програм для обробки природної мови. Останнім часом методи глибокого навчання на прикладах цієї проблеми досягли найсучасніших результатів. Було продемонстровано, що моделі глибокого навчання здатні досягати оптимальних результатів у сфері проблем генерації підписів. Замість того, щоб вимагати складної підготовки даних або конвеєра спеціально розроблених моделей, можна визначити одну наскрізну модель, щоб передбачити підпис за фотографією. нові алгоритми які показують себе найкраще для конкретних дата сетів.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Загальний аналіз галузі

Для більшості зображень люди можуть відносно легко підготувати стислий опис у формі речення. Такі описи можуть визначити найцікавіші об’єкти, що вони роблять і де це відбувається. Ці описи є інформативними для людини, тому що вони у формі речень. Якщо порівнювати методи Image captioning до того як працює мозок людини, то наш мозок має повноцінну когнітивну систему. Поки є можливість отримувати образи зображені на картинці, мозок буде аналізувати їх зміст і обробляти цю інформацію, щоб зробити власні висновки що саме зображено на картинці. Але коли комп’ютер генерує підписи до зображень, виникає потрібність надати комп’ютеру щось схожого на здатність когнітивного розуміння змісту зображення [2].

Image Captioning – це напрям досліджень, які вивчають перетворення візуальних образів у інші форми інформації, які можуть сприйматися людиною. Зазвичай базова модель таких методів складається з двох частин або фаз. Перша частина – це вилучення інформації щодо особливостей зображення, яка переважно полягає у вилучення інформації про об’єкти та локацію їх розташування у зображені. У другій фазі виконується аналіз семантичної інформації зображення та об’єднання її за допомогою спеціальних функцій з інформацією щодо особливостей, яку було отримано під час першої фази. Традиційні, або як їх ще кличуть евристичні, методи програмування не можуть реалізувати цю функцію. З одного боку логіка, яку слід розглянути, занадто складна, а програма занадто велика, з іншого боку, обмеження традиційних програм занадто жорсткі для досягнення очікуваного ефекту. Алгоритми Image Captioning можно поділити на дві основні групи, один заснований на методі шаблонів(template methods), другий заснований на структурі кодера та декодера(encoder decoder structure).

Початкові алгоритми Image Captioning використали у своєї більшості метод шаблонів. Цей метод спочатку витягує серію ключових даних про

особливості зображення, такі як ключові об’єкти, спеціальні атрибути, інформацію щодо розташування об’єктів, за допомогою різних типів класифікаторів, таких як SVM [3], а потім використовує лексичну модель або інші специфічні шаблони для перетворення отриманої інформації про особливості в текст опису. Так наприклад алгоритм запропонований Алі Фархаді та його колегами [4], представляв множину значень реченнями які складалися з трьох складових: об’єкт, дія, сцена. (Рисунок 2.1). Такі триплети мають назву MRF. Цей триплет є достатнім, щоб надати загальну ідею, про те що коїться на зображені. Зазвичай це співпадають з тим, про що людина говорить як тільки побачить світлину, це можна порівняти з найкоротшим резюме. Для кожного слота в триплеті існує дискретний набір можливих значень.

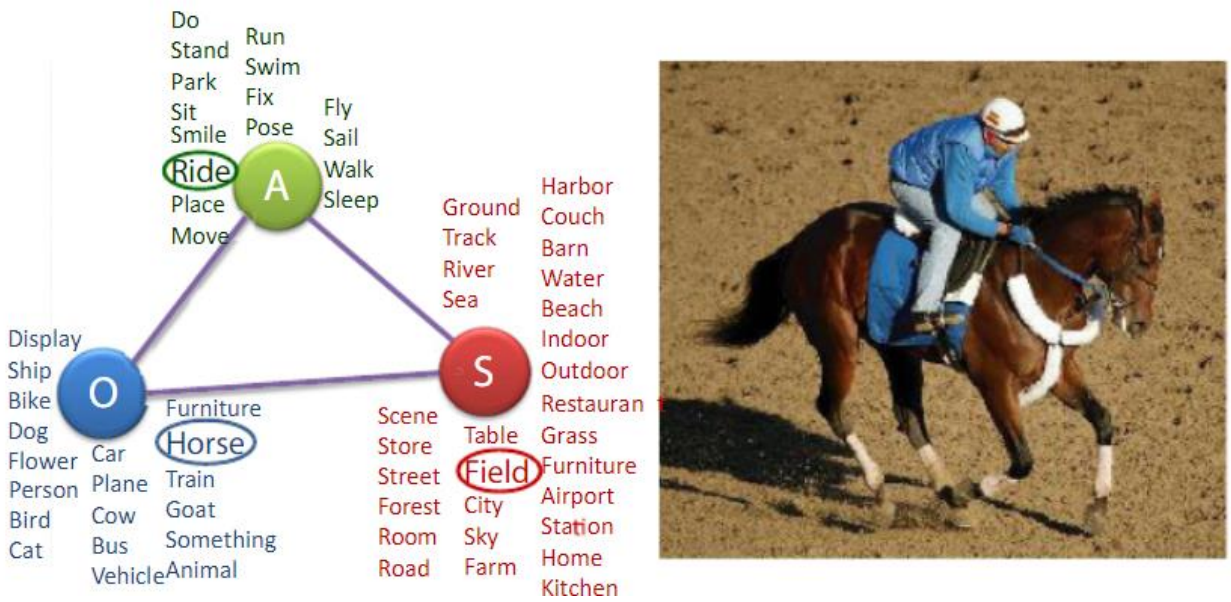

Рисунок 1.1 – Приклад складання речення у шаблонних методах

«O» означає вузол для об’єкта, «A» для дії, а «S» для сцени. Навчання включає встановлення ваг для потенціалів вузла та краю, а висновок — це пошук найкращих трійок з урахуванням потенціалів.

Така модель передбачає, що між простором речень і простором образів є простір значень, який оцінюють подібність між реченням і зображенням шляхом

(а) відображення кожного з проміжком значень, а потім (б) порівняння результатів.

Потенціали вузлів обчислюються за допомогою лінійної комбінації балів кількох детекторів і класифікаторів. Крайові потенціали оцінюються за частотами. Можливі значення для кожного вузла, кожен з яких знаходиться у простору станів достатнього розміру, записані на зображенні.

Шаблонні методи зазвичай вчяться передбачати трійки для зображень дискримінаційно. Для цього потрібно мати набір даних зображень(dataset), позначених їх значеннями триплетів. Потенціали обчислюються як лінійні комбінації функцій для виявлення ознак(feature functions). Це ставить проблему навчання як пошук найкращого набору ваг для лінійних комбінацій функцій властивостей, щоб триплет основної істини оцінили вище, ніж будь-які інші триплети. Висновок включає знаходження наступної функції:

$$argmax_y \omega^T \Phi(x, y),$$

де y — мітка триплету;

ω — вивчені ваги;

Ф – це потенціал функції.

Методи які працюють за структурою кодера та декодера є більш гнучкими. У кодері нейронна мережа згортки використовується для вилучення та векторизації інформації про основні характеристики зображення. Декодер переважно використовує рекурентну нейронну мережу, яка об'єднує векторизоване зображення функції із семантичною інформацією для створення опису змісту зображення.

З появою механізмів імітуючих когнітивну увагу, дослідники почали використовувати такі механізм для вирішення проблеми Image Captioning. Було запропоновано два механізми уваги та застосовування їх до різних типів областей зображення [3]. Механізм м’якої уваги фокусується на всіх областях

зображення, і значення ваги кожної області різне. Чим вище значення, тим важливіше регіон у прогнозі. Контекстний вектор, який буде використано у результаті, отримують розподілом ваги кожного регіону. На відміну від механізму м'якої уваги, механізм жорсткої уваги фокусується лише на певній області. Щоб збільшити гнучкість зазвичай області уваги вибираються випадково, а потім вже розраховується правдоподібність обраних областей.

Модель яка використовує механізм семантичної уваги показана на Рисунку 2.2. Ця модель поєднує в собі підходи для опису характеристик зображень «знизу вгору» (bottom-up) і «зверху вниз»(top-down). Мережа використовує механізми уваги щоб гнучко й точно відповідно до важливих особливостей зображень отримувати атрибути, а також брати важливу семантичну інформацію як центр уваги.

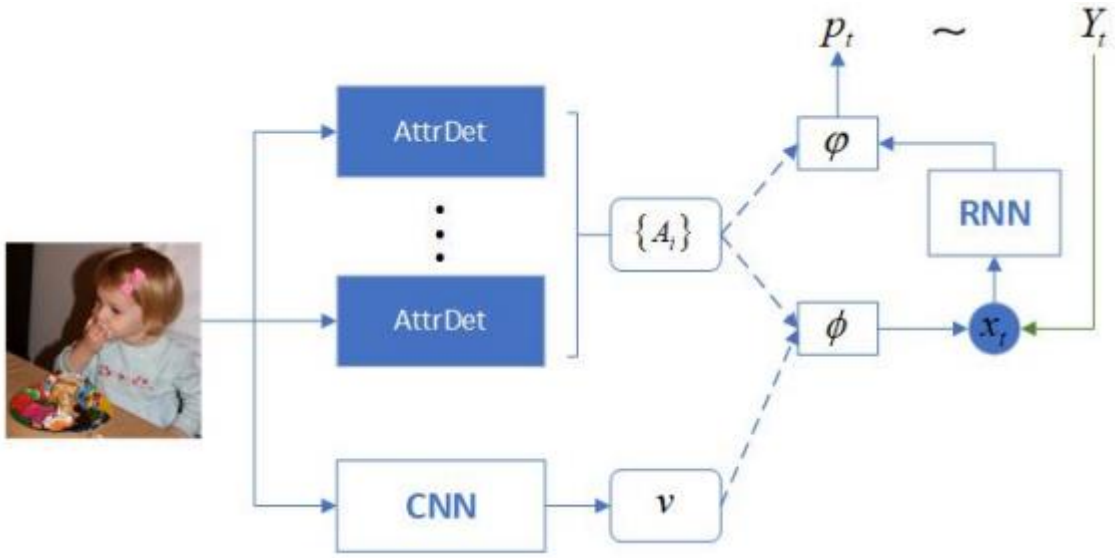

Рисунок 1.2 – Структура роботи одного з алгоритмів за принципом кодеру та декордеру

По-перше, інформація про характеристики зображення, оброблена загортковою нейронною мережею (CNN або convolution neural network) [5], використовується для того, щоб повторювана нейронна мережа (RNN або recurrent neural network) містила інформацію про візуальні характеристики; потім

для виявлення візуальних об'єктів або атрибутів використовується детектор атрибутів “AttrDet”; нарешті, вони об’єднуються, щоб отримати остаточну послідовність опису зображення за допомогою повторюваної нейронної мережі.

Рекуррентні нейронні мережі, довготривала короткочасна пам'ять і рекуррентні нейронні мережі з гальмуванням, зокрема, міцно зарекомендували себе як сучасні підходи в моделюванні послідовності та проблемах трансдукції, таких як мовне моделювання та машинний переклад. З тих пір численні зусилля продовжували розширювати межі рекурентних мовних моделей та архітектур кодувальників-декодерів. Рекурентні моделі, як правило, розраховують на фактори позиції символів вхідних і вихідних послідовностей. Вирівнюючи позиції до кроків у часі обчислення, вони генерують послідовність прихованих станів h_t, як функцію попереднього прихованого стану h_{t-1} та вхідних даних для позиції t. Ця послідовна природа виключає розпаралелювання в навчальних прикладах, що стає критичним при більшій довжині послідовності, оскільки обмеження пам’яті обмежують групування прикладів. Нещодавні роботи дозволили досягти значного покращення обчислювальної ефективності за допомогою трюків факторізації та умовного обчислення, а також покращення продуктивності моделі у випадку останнього. Однак фундаментальне обмеження послідовного обчислення залишається.

Механізми уваги стали невід'ємною частиною переконливого моделювання послідовностей і моделей трансдукції в різних задачах, дозволяючи моделювати залежності без урахування їх відстані у вхідних або вихідних послідовностях. Однак у всіх, за винятком кількох випадках, такі механізми уваги використовуються в поєднанні з рекурентною мережею. Архітектура моделі Transformer уникає повторюваності та натомість повністю покладається на механізм уваги для малювання глобальних залежностей між входом і виходом.

1.2 Виявлення проблем

Комп'ютерний зір в області обробки зображень досяг значних успіхів, як-от класифікація зображень та виявлення об'єктів. Цей напрям у дав розвиток більш складному але не менш важному напряму Image Captioning. Автоматична генерація повних і природних описів зображень має великі потенційні наслідки, такі як:

– створення заголовків, доданих до зображень новин;

– описи, пов’язані з медичними зображеннями;

– пошук зображень на основі тексту;

– інформація, доступна для сліпих користувачів;

– покращення взаємодія людини і робота.

Ці застосування в підписах зображень мають важливе теоретичне та практичне значення для дослідження. Тому створення підписів до зображень є більш складним, але значущим завданням в епоху штучного інтелекту.

Наразі або виділяють два групи методів Image Captioning: «шаблонні» методи і методи засновані на патерні кодера та декорера. Шаблонні методи занадто сильно покладаються на фіксовані правила та отриману конкретну інформацію про зображення, і різні правила безпосередньо впливатимуть на згенеровані оператори та речення. Вміст опису зображення, отриманий такими методами, занадто простий і не має універсальності, що часто не дозволяє досягти бажаного ефекту. Саме тому все більше популярність отримують методи засновані на патерні кодера та декодера. Такі методи генерації підписів зображень є більш гнучкими у передбаченні і дозволяють отримувати більш детальні речення про те що зображено на картинці.

Оцінити генерацію речень дуже складно, тому що речення плавні, і досить різні речення можуть описувати одні й ті ж явища. Що ще гірше, синекдоха (наприклад, заміна «тварина» на «кішка» або «велосипед» на «автотранспорт») і загальне багатство словникового запасу означають, що багато різних слів цілком законно можна використовувати для опису однієї і тієї ж картини. Але зараз вже

існують і активно вивчаються алгоритми, які дозволяють отримати state-of-the-art продуктивність. Різні алгоритми залежачи від датасету, мають кожен свій різний результат за метриками BLEU and METEOR. І залежачи від того чи іншого датасету, який використається для вивчання, можна підібрати алгоритм який показує найкращий результат для конкретного датасету.

1.3 Постановка задачі

Основною задачею цієї роботи є дослідження методів генерації підписі до зображень (Image Captioning). Необхідно дослідити актуальні сучасні моделі, які вже використовуються для даної задачі, виявити популярні метрики для оцінювання результативності підписів, дослідити існуючі датасети для моделей Image Captioning, обрати одну з моделей, яка дозволяє отримувати одні з кращих оцінок згідно з обраними метриками, реалізувати обрану модель та навчити її на обраному датасеті. Після навчання обраної моделі необхідно порівняти отримані оцінки згідно обраних метрик та порівняти результати з іншими і існуючими моделями. У роботі пропонується використовувати мережу Faster R-CNN для виявлення об’єктів на зображенні описану в пункті 3.2, зумисно з моделлю трансформера об'єктних відносин описану в пункті 3.3.

2 МЕТРИКИ ОЦІНКИ ЕФЕКТИВНОСТІ АЛГОРИТМУ

2.1 Перелік існуючих метрик

Людська оцінка машинного перекладу (MT) враховує багато аспектів перекладу, включаючи адекватність, точність і плавність перекладу. Але велика частина технік, які використовуються при такої оцінці є досить дорогими. Більше того, для їх завершення можуть знадобитися тижні або місяці, що є неприпустимо при розробці систем машинного перекладу, під час того якої потрібно стежити за ефектом щоденних змін у своїх системах, щоб відсівати погані методи та правити помилки.

Тому для оцінювання машинного перекладу були розроблені техніки, які також використовуються при оцінці результатів Image Captioning. Не дивлячись на те як активно розвивається напрямок Image Captioning і як часто з’являються нові алгоритми, розробка показників оцінки для підписів зображень, залишається незмінною останніми роками. Найчастіше для алгоритмів генерації підписів використовують наступні показники:

– BLEU;
– METEOR;
– ROUGE;
– SPICE;
– CIDER.

Всі ці показники розраховують жорстке вирівнювання між запропонованими підписами і базовими правдивими підписами. Щоб зрозуміти, від чого залежить оцінка згенерованого підпису розглянемо, що саме враховує кожна оцінка і як саме вони працюють.

2.2 Огляд роботи метрики BLEU

Як було зазначено раніше, ця метрика була розроблена для оцінювання машинних перекладів. Основним компонентом BLEU є n-грамна точність: частка узгоджених n-грам від загальної кількості n-грам в оцінюваному перекладі.

Точність розраховується окремо для кожного порядку n-грам, а точність об’єднується за допомогою геометричного усереднення. У цієй метрці рахується середнє геометричне змінених балів точності тестового корпусу, а потім результат множиться на експоненційний коефіцієнт погашення короткості. Згортання регістру є єдиною нормалізацією тексту, яка виконується перед обчисленням точності.

Цю оцінку рахують за наступною формулою:

$$BLEU = \frac{\sum_i \sum_k \min\{h_k(c_i), max_j\ h(s_{ij})\ \}}{\sum_i \sum_k h_k(c_i)},$$

де k,j є індексами N-грамного маркеру і тестового кейсу відповідно;

h — кількість зустрічей токена;

C і S — запропоновані і правдиві підписи відповідно.

Отже, $h_k(c_i)$ означає кількість входжень k-го слова в і-у підпис кандидата. З цієї формули метричного виразу ми бачимо, пропорцію яка описує скільки згенерованих слів збігалося з написом які беруться, як істинні з усіх N-слів згенерованого речення. Щоб зменшити прив’язку результата до конкретного речення, у формулі використається таке істинне речення, у якому слова з згенерованого речення зустрічаються найчастіше (зазвичай для кожного зображення є 5 речень які використаються як істина). Щоб уникнути ситуацій, коли одне повторюване слово набирає повну оцінку, береться мінімальна кількість зустрічей слова між згенерованому реченням та в тим, яке використовується як істина. Після повторення всіх N-слів у згенерованому реченні всі оцінки отриманих N-слів(N дорівнює 1,2,3 або 4) потім підсумовуються, щоб отримати остаточну оцінку речення кандидата з додаванням короткого терміну покарання. Зрештою ми отримуємо бали BLEU.

Показник BLEU коливається від 0 до 1. Дуже небагато перекладів отримають оцінку 1, якщо вони не ідентичні еталонному перекладу. З цієї

причини навіть речення сформульовано людиною не обов’язково отримає 1 бал. Важливо зазначити, що чим більше довідкових перекладів на речення, тим вище оцінка.

2.3 Огляд роботи метрики ROUGE

Цей вигляд метрики спочатку був запропонований для текстових підсумкових завдань, він є дуже схожим на метрику BLEU, але надає більше уваги балансу між влучністю та повнотою згенерованих речень, в той час коли метрика BLEU враховує лише повноту згенерованих речень, а влучність залишає без уваги. Існує кілька варіацій цієї оцінки: ROUGE-L, ROUGE-N, ROUGE-S, ROUGE-W, ROUGE-SU. Саме для Image Captioning алгоритмів використовують ROUGE-L варіацію, яка показує найдовшу загальну послідовність. Ця оцінка рахується за наступними формулами:

$$R_{lcs} = \frac{\mathrm{LCS(X,Y)}}{m},$$

$$P_{lcs} = \frac{\mathrm{LCS(X,Y)}}{n},$$

$$\beta = \frac{R_{lcs}}{P_{lcs}},$$

$$\mathrm{ROUGE_L} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}},$$

де LCS(X,Y) — це довжина найдовшої спільної підпослідовності множин X і Y;

m — довжина множини X;

n — довжина множини Y.

Таким чином ця оцінка максимальна, і дорівнює одиниці коли X = Y, і мінімальна(дорівнює 0) при LCS(X,Y) = 0, тобто коли в двох реченнях немає спільних слів. Метрика ROUGE-L відповідає кільком теоретичним критеріям точності вимірювання, що включають більше одного фактора. У цьому випадку складовими факторами є повнота та влучність на основі LCS. Таким чином цю

оцінку використовують для порівняння схожості двох речень і показали і у деяких випадках [6] вона показує себе більш ефективною ніж метрика BLEU.

2.3 Огляд роботи метрики METIOR

Ця метрика при розробці теж була призначена для машинного перекладу, яку можна описати як іншу покращену версію оригінальної BLEU метрики. Ця метрика у своїй цілі ставить наступні покращення порівняно з BLEU метрикою:

- більша повнота(recall) оцінки;
- використання N-грам більш вищого порядку. N-грами вищого порядку використовуються в BLEU як непрямий показник рівня граматичної правильності перекладу. У METIOR метриці явний міра рівня граматичності (або порядку слів) сприяє кращому розрахунку граматичності як фактора машиного перекладу і приводе до кращої кореляції з людськими судженнями про якість перекладу.
- відсутність чіткої прив’язки слів між запропонованим реченням і реченням яке вважається істинною. Підрахунок N-грам не вимагає явного співставлення між словом, але це може призвести до підрахунку неправильних «збігів», особливо для загальних функціональних слів.
- використання геометричного усереднення N-грамів. Геометричне усереднення призводить до оцінки «нуль», коли один із компонентних балів n-грам дорівнює нулю. Отже, оцінки BLEU на рівні речення (або сегмента) можуть бути безглуздими, що не дуже зручно при використанні цієї метрики для оцінки речень згенерованих Image Captionging моделями. Хоча BLEU планувалося використовувати лише для агрегованих підрахунків протягом усього тестового набору (а не на рівні речення), результати на рівні пропозиції можуть бути корисними індикаторами якості метрики. Так як мерика METIOR використовує рівновагову арифметичну середню оцінку n-грам, вона має кращу кореляцію з судженнями людини.

На основі F показника попередньої метрики ROUGE, тут новим змінною та складовою є показник покарання за частину речення:

$$F_{mean} = \frac{10PR}{R + 9P},$$

де P — точність першої уніграми, обчислюється як відношення кількості уніграм у системному перекладі до загальної кількості уніграм у системному перекладі.

Аналогічно, відкликання уніграм (R) обчислюється як відношення кількості уніграм у системному перекладі до загальної кількості уніграм у еталонному перекладі. Влучність, повнота та F показник засновані на збігах уніграм. Усі уніграми в згенерованому реченні, які відображаються в уніграми в істинному реченні, групуються в найменшу можливу кількість фрагментів, щоб уніграми в кожній частині були в сусідніх позиціях у згенерованому реченні, а також відображалися в уніграми, які знаходяться в сусідніх позиціях у реченні яке може сприйматися як істинне. Таким чином, чим довші n-грами, тим менше фрагментів, і в крайньому випадку, коли весь системний рядок перекладу відповідає опорному перекладу, залишається лише один фрагмент. З іншого боку, якщо немає біграмних або довших збігів, то кількість фрагментів збігається з кількостю збігів уніграм. Показник покарання (Penalty) розраховується за допомогою наступної формули:

$$\text{Penalty} = 0.5 * (\frac{\#chunks}{\#\text{unigrams matched}})^3,$$

Де chunks — кількість фрагментів;

unigrams matched — кількість збігів уніграм.

Нарешті, оцінка METEOR для даного вирівнювання обчислюється наступним чином:

$$\text{Score} = F_{mean} * (1 - \text{Penalty}),$$

Цей алгоритм автоматично знаходить непересічні схожі пари фрагменти між реченням кандидатом і реченням істиною [7]. Якщо речення кандидат строго збігається з еталоном, то цей показник досягає свого мінімуму, і в цьому випадку бал METEOR досягає максимального рівня.

2.4 Огляд роботи метрики CIDER

На відміну від раніше зазначених метрик, цей метод був заснований саме для оцінювання речень згенерованих для опису малюнків. Основним механізмом, вбудованим у CIDER, є термін зважування tf-idf [8], який має широкий спектр застосування у векторній семантиці. Основна ідея цього терміна випливає з припущення, що частота зустрічальності слова в певному корпусі може бути значущим показником характеру чи семантичного значення слова. Деякі слова часто зустрічаються майже в кожному корпусі («і», «те» тощо). Отже, коли ми розглядаємо значення цілого речення, нам потрібно зважити різні слова на значення для цього питання.

Мета цієї метрики автоматично оцінити як добре для зображення I_i, відповідає речення-кандидат c_i відповідає консенсусу набору описів зображень S_i={$S_{i1}, \ldots, S_{im}$}. Усі слова в реченнях (як кандидати, так і посилання) спочатку зіставляються з їх основними або кореневими формами. У метриці кожне речення представляється, використовуючи набір n-грам, присутніх у ньому, де n-грама ω_k — це набір одного або кількох упорядкованих слів.

Інтуїтивно, міра консенсусу буде кодувати, як часто n-грами в реченні-кандидаті присутні в реченнях-істини, або як їх ще називають реченнях-посиланнях. Аналогічно, n-грами, яких немає в реченнях-посиланнях, не повинні

бути в реченні-кандидаті. Нарешті, n-грамам, які зазвичай зустрічаються на всіх зображеннях у наборі даних, слід надавати меншу вагу, оскільки вони, ймовірно, будуть менш інформативними. Щоб закодувати цю інтуїцію, ми виконуємо зважування частоти інверсного документа Term Frequency (TF-IDF) для кожного n-грама [9]. Кількість разів, коли n-грама ω_k зустрічається в опорному реченні S_{ij}, позначається $h_k(s_{ij})$ або $h_k(c_i)$ для пропозиції c_i. Ми обчислюємо зважування TF-IDF $g_k(s_{ij})$ для кожного n-грама ω_k, використовуючи:

$$g_k(s_{ij}) = \frac{h_k(c_i)}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min\left(1, \sum_q h_k(s_{pq})\right)}\right),$$

де Ω — словниковий запас усіх n-грамів;

I — множина усіх зображень у наборі даних.

Перший доданок вимірює TF кожного n-грама ω_k, а другий доданок вимірює рідкість ω_k, використовуючи його IDF. Інтуїтивно, TF надає більшу вагу n-грамам, які часто зустрічаються в опорному реченні, що описує зображення, тоді як IDF зменшує вагу n-грам, які зазвичай зустрічаються в усіх зображеннях у наборі даних. Тобто IDF забезпечує міру помітності слів, знижуючи популярні слова, які, ймовірно, будуть менш візуально інформативними. IDF обчислюється за допомогою логарифма кількості зображень у наборі даних |I| поділено на кількість зображень, для яких ω_k зустрічається в будь-якому з його опорних речень.

Оцінка CIDERn для n-грамів довжини n обчислюється з використанням середньої подібності косинуса між реченням-кандидатом і реченням-посиланням, що враховує як влучність, так і повноту:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g^n}(c_i) \cdot \boldsymbol{g^n}(s_{ij})}{\|\boldsymbol{g^n}(c_i)\| \|\boldsymbol{g^n}(s_{ij})\|}$$

де $g^n(c_i)$ — вектор, утворений $g_k(c_i)$, що відповідає всім n-грамам довжини n $||g^n(c_i)||$ — величина вектора $g^n(c_i)$. Аналогічно для $g^n(S_{ij})$.

Для захоплення граматичних властивостей, а також розширеної семантики, використовуються n-грами вищого порядку (довші). Оцінки з n-грамів різної довжини поєднуються наступним чином:

$$CIDER(c_i, S_i) = \sum_{n=1}^{N} \omega_n \, CIDER(c_i, S_i),$$

де імпірично було виявлено, що рівномірні ваги ω_n= 1/N працюють найкраще.

2.5 Огляд роботи метрики SPICE

Замість того, щоб розглядати відповідність з N-грамами в реченні, як і попередні показники, SPICE творчо використовує кортежі слів і обчислює перетин між згенерованими реченнями та основними істинними реченнями. Ще однією особливістю метрики SPICE є графічне представлення речень. Показник SPICE [10] охоплює більше інформації, ніж токен N-грам. Процес зіставлення здійснюється шляхом порівняння двох наборів кортежів текстів і графів зображень. Приклад такого графа зображено на малюнку 2.1. Однак продуктивність цієї метрики багато в чому залежить від точності аналізу. Він віддає перевагу довшим реченням і схильний втрачати структурну інформацію речень.

SPICE вимірює, наскільки добре генератори підписів виявляють об'єкти, атрибути та зв'язки між ними. Потенційна проблема полягає в тому, що алгоритмами генерації можуть обдурити показник, створюючи підписи, які представляють лише об’єкти, атрибути та відносини, ігноруючи інші важливі аспекти граматики та синтаксису. Оскільки SPICE нехтує плавністю, як і у випадку з метриками n-грам, він неявно припускає, що підписи правильно

сформовані. Якщо це припущення не відповідає дійсності в конкретному додатку, в оцінку можна включити показник плавності, такий як власна інформація (surprisal) [11]. Однак за замовчуванням до цієї метрики не включені жодни коригування плавності, оскільки концептуально перевагу віддана простішим і легшим інтерпретованим показникам.

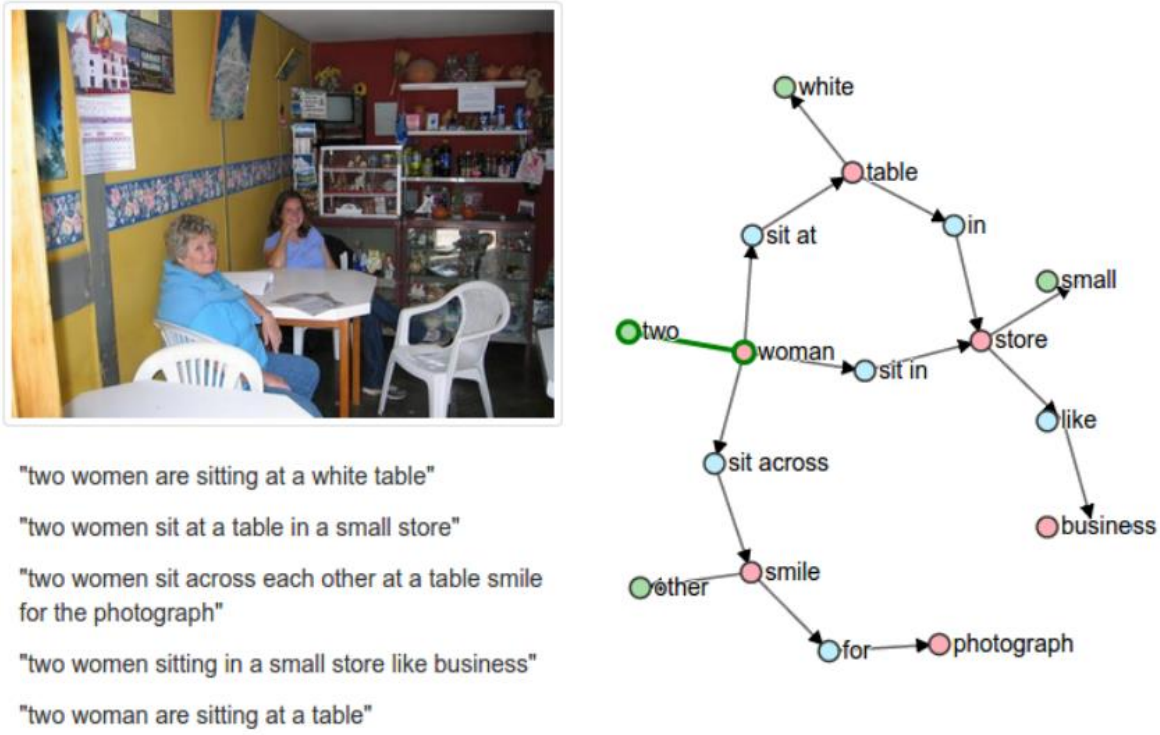

Рисунок 2.1 — Приклад графіка сцени (праворуч), який використається в SPICE метриці, розібраного з набору підписів опорних зображень (ліворуч).

Щоб якомога точніше моделювати людське судження для певного речення яке розглядається, найкраще використовувати ретельно налаштований ансамбль показників, включаючи SPICE, що фіксує різні виміри правильності. Саме тому при оцінці роботи моделей Image Captioning використовують одразу кілька метрик, і всі метрики які було розглянуто у цьому розділі є найчастіше використовуваними.

3 ВИБІР ТА ОПИС РОБОТИ АЛГОРИТМУ

3.1 Вибір та обґрунтування алгоритму

Багато з нейронних моделей для підписів зображень кодували візуальну інформацію за допомогою одного вектора ознак, що представляє зображення в цілому, і, отже, не використовували інформацію про об’єкти та їх просторові відносини. Винятком є такі моделі як наприклад Deep visual-semantic alignments [12], у якої витягають ознаки з кількох областей зображення на основі детектора об’єктів R-CNN [13] щоб створювати окремі підписи для регіонів. У цій моделі для кожного регіону створюється окремий підпис, проте просторовий зв’язок між виявленими об’єктами не моделюється. Це також стосується і моделей пов’язаних з щільними підписами (dense captioning) [14], які представляють наскрізний підхід для отримання підписів до зображення, що стосуються різних регіонів зображення. В таких моделях просторова асоціація була здійснена шляхом застосування повністю згорткової нейронної мережі до зображення та створення карт просторових відповідей для цільових слів, але відносини між просторовими регіонами не були взяти до уваги.

З іншої сторони існують сімейство Image Captioning алгоритмів, що ґрунтуються на увазі, які прагнуть обґрунтувати слова в передбачуваному підписі до регіонів зображення. Оскільки візуальна увага часто привертається до вищих згорткових шарів CNN, просторова локалізація обмежена і часто не має семантичного значення. Цю проблему вирішують моделі засновані на обмеженні типових моделей уваги шляхом поєднання моделі уваги «знизу вгору» з LSTM «зверху вниз». Можель уваги «знизу вгору» діє на згорткові ознаки, об’єднані середнім об’ємом, отримані із запропонованих регіонів інтересу для детектора об’єктів Faster R-CNN [14]. Модель LSTM «зверху вниз» — це двошаровий LSTM, у якому перший шар діє як модель візуальної уваги, який звертає увагу на відповідні виявлення для поточного маркера, а другий рівень — це мовний LSTM, яка генерує наступний маркер. За допомогою можливо достигнути найсучасніших характеристики як візуальних відповідей на запитання, так і

підписів до зображень, вказуючи на переваги поєднання функцій, отриманих від виявлення об’єктів, із візуальною увагою. Моделі які використовують геометричну увагу для виявлення об'єктів використовували координати та розміри обмежувальної рамки, щоб зробити висновок про важливість взаємозв’язку пар об’єктів, припускаючи, що якщо дві обмежувальні прямокутники ближчі та більш подібні один до одного, то їх зв’язок сильніший.

Одна з відомих сучасних моделей, яка дотримується вищезгаданої парадигми отримання характеристик зображення за допомогою детектора об’єктів та створення підписів використовує модель LSTM уваги. У цьому констексті доречно згадати двох-графовову згорктову мережу (Two Graph Convolutional Networks) засновану на графі семантичних зв'язків і графі просторових зв'язків, яка класифікує зв'язок між двома блоками у 11 класів, таких як «всередині», «покриття» або «перекриття». Крім того, у такій моделі використовувують семантичний графік сцени (тобто графік об’єктів, їхніх зв’язків та їхніх атрибутів) автокодер (autoencoder) у тексті підпису, для вбудовування мовного індуктивного упередження в словник, який використовується спільно з графіком сцени зображення. Хоча ця модель може вивчати типові просторові зв’язки, знайдені в тексті, вона за своєю суттю не здатна охопити візуальну геометрію, характерну для даного зображення.

На противагу моделі заснованій на двох-графововій згорктовій мережі існує інший підхід представлений у алгоритму трансформатора об'єктних відносин, який безпосередньо використовує співвідношення розмірів і різницю координат рамки, неявно кодуючи та узагальнюючи вищезгадані відносини. Під час навчання такої моделі важливу роль грає самокритичного навчання з підкріпленням для генерації речень, що описують малюнки. Модель алгоритму трансформатора об'єктних відносин є покращеною версію архітектури Transformer [10], або Standart Transformer, яка була створена для покращення продуктивності таких завдань як: переклад, генерація тексту та розуміння мови. У моделі Standart Transformer використовують глобальну інформації про все

зображення, так само як і інформації з окремих рівномірних 8x8 частин зображення. Вектори ознак послідовно подавали в кодер частину моделі. Архітектуру моделі Transformer зображена на рисунку 3.1.

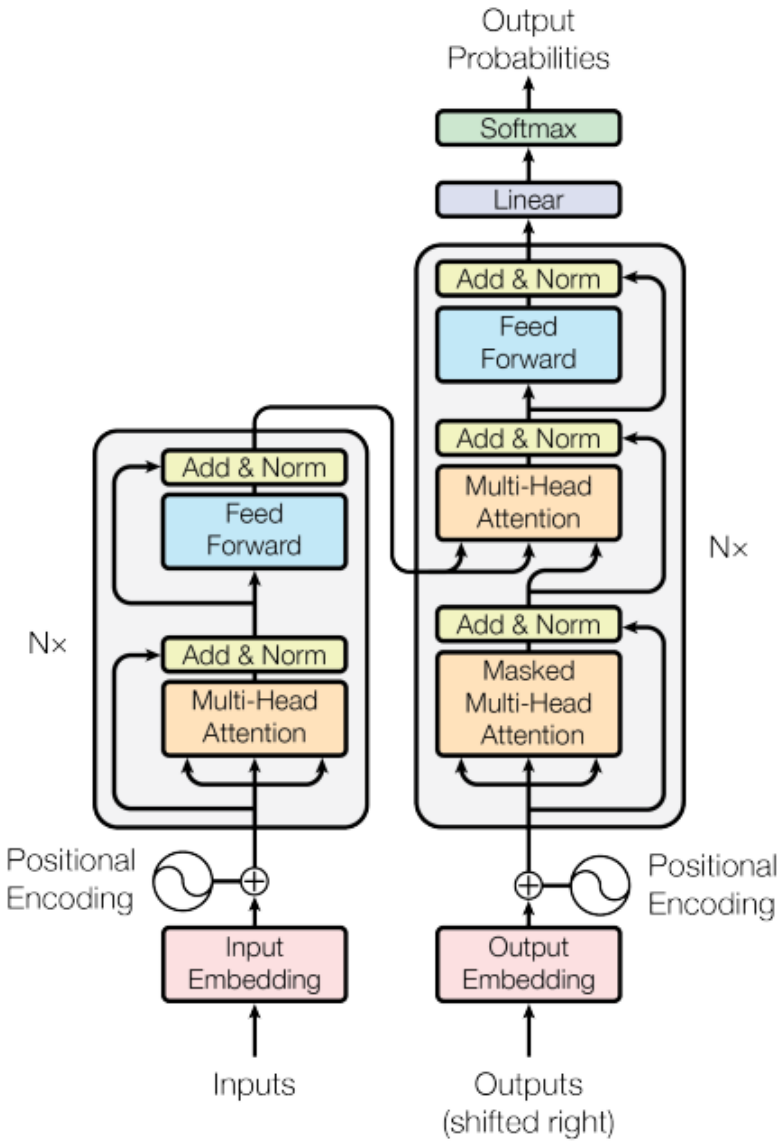

Рисунок 3.1 – Архітектура моделі Трансформера

Архітектура Transformer особливо добре підходить як «знизу вгору» візуальний кодер для підписів до зображень, оскільки вона не має поняття порядку для своїх вхідних даних, на відміну від RNN. Однак ця модель може успішно моделювати послідовні дані з використанням позиційного кодування,

яке застосовується до декодованих маркерів у тексті підпису в покращеній моделі трансформатора об'єктних відносин (Object Relation Tranformer). Замість того, щоб кодувати порядок для трьох об’єктів, трансформатор об'єктних відносин прагне закодувати, як два об’єкти просторово пов’язані один з одним, і відповідно зважувати їх, що дозволяє покращити оцінки отриманими для навченої моделі з використанням основних метрік, таких як CIDER або METEOR.

3.2 Виявлення об'єктів на зображенні

Натхнені нейронним машинним перекладом, більшість звичайних систем підписання зображень використовують структуру кодер-декодер, у якій вхідне зображення кодується у проміжне представлення інформації, що міститься в зображенні, а потім декодується в описову текстову послідовність. Це кодування може складатися з одного вихідного вектора ознак CNN [11] або кількох візуальних ознак, отриманих з різних регіонів зображення. В останньому випадку вибірка областей може бути рівномірна [12] або керуватися детектором об’єктів, який, як було показано, дає покращені характеристики.

Хоча ці кодери, засновані на детектуванні, представляють найсучасніший рівень техніки, наразі вони не використовують інформацію про просторові відносини між виявленими об’єктами, наприклад відносне положення та розмір. Однак ця інформація часто може бути важливою для розуміння вмісту зображення, і вона використовується людьми, коли справа йде про фізичний світ. Відносна позиція, наприклад, може допомогти відрізнити «дівчину, яка їде на коні» від «дівчинки, що стоїть біля коня». Аналогічно, відносний розмір може допомогти розрізнити «жінку, яка грає на гітарі» та «жінку, яка грає на укелеле». Було доказано, що включення просторових зв'язків покращує продуктивність самого виявлення об'єктів. Крім того, у кодувальниках машинного перекладу позиційні відносини часто кодуються, зокрема у випадку Transformer, який використовує архітектури кодера засновану на механізму семантичної уваги.

Як базову CNN для виявлення об’єктів та вилучення ознак у обраної моделі використовується Faster R-CNN [6] з ResNet-101. Використовуючи проміжні карти характеристик з ResNet-101 як вхідні дані, мережа регіональних пропозицій (RPN) генерує обмежувальні рамки для пропозицій об’єктів. Використовуючи немаксимальне стримування, обмежувальні прямокутники, що перекриваються, з перетином над об’єднанням (IoU), що перевищує поріг 0,7, відкидаються. Потім шар об’єднання регіону інтересу (RoI) використовується для перетворення всіх обмежувальних прямокутників, що залишилися, на той самий просторовий розмір (наприклад, 14 × 14 × 2048). Додаткові шари CNN застосовуються для прогнозування міток класів і уточнення рамки для кожної пропозиції прямокутнику. Далі ми відкидаємо всі обмежувальні прямокутники, у яких ймовірність передбачення класу нижче порогу 0,2. Нарешті, ми застосовуємо середнє об’єднання для просторового виміру, щоб створити 2048-вимірний вектор ознак для кожного обмежувального прямокутника об’єкта. Ці вектори ознак потім використовуються як вхідні дані для моделі трансформатора.

Найсучасніші мережі виявлення об’єктів залежать від алгоритмів пропозицій регіону для гіпотези розташування об’єктів. У моделі Fast R-CNN, час роботи необхідний для виявлення, а відповідно і продуктивність, було покращено, тому що обчислення пропозицій регіону було виявлено як вузьке місце. У цій моделі використовується мережа регіональних пропозицій (RPN), яка використовує згорткові функції повного зображення спільно з мережею виявлення, що дозволяє пропозиції регіонів майже безкоштовні у сенсі технічних ресурсів комп’ютеру. RPN — це повністю згорткова мережа, яка одночасно прогнозує межі об’єкта та оцінки об’єктності в кожній позиції. RPN навчаються з початку до кінця для створення високоякісних пропозицій регіонів, які використовуються Fast R-CNN для виявлення об’єктів. За допомогою простої альтернативної оптимізації RPN і Fast R-CNN можна навчити спільно використовувати згорткові функції.

Мережа регіональних пропозицій (RPN) приймає зображення (будь-якого розміру) як вхідні дані та виводить набір пропозицій прямокутних об’єктів, кожна з яких має оцінку об’єктності. Щоб генерувати пропозиції регіонів, ми переміщаємо невелику мережу через карту згорткових об’єктів, виведену останнім спільним згортковим шаром. Ця мережа повністю підключена до n × n просторового вікна карти об’єктів вхідного карти згорткових характеристик. Спрощена схема роботи мережа регіональних пропозицій в одному місці зображено на рисунку 3.2.

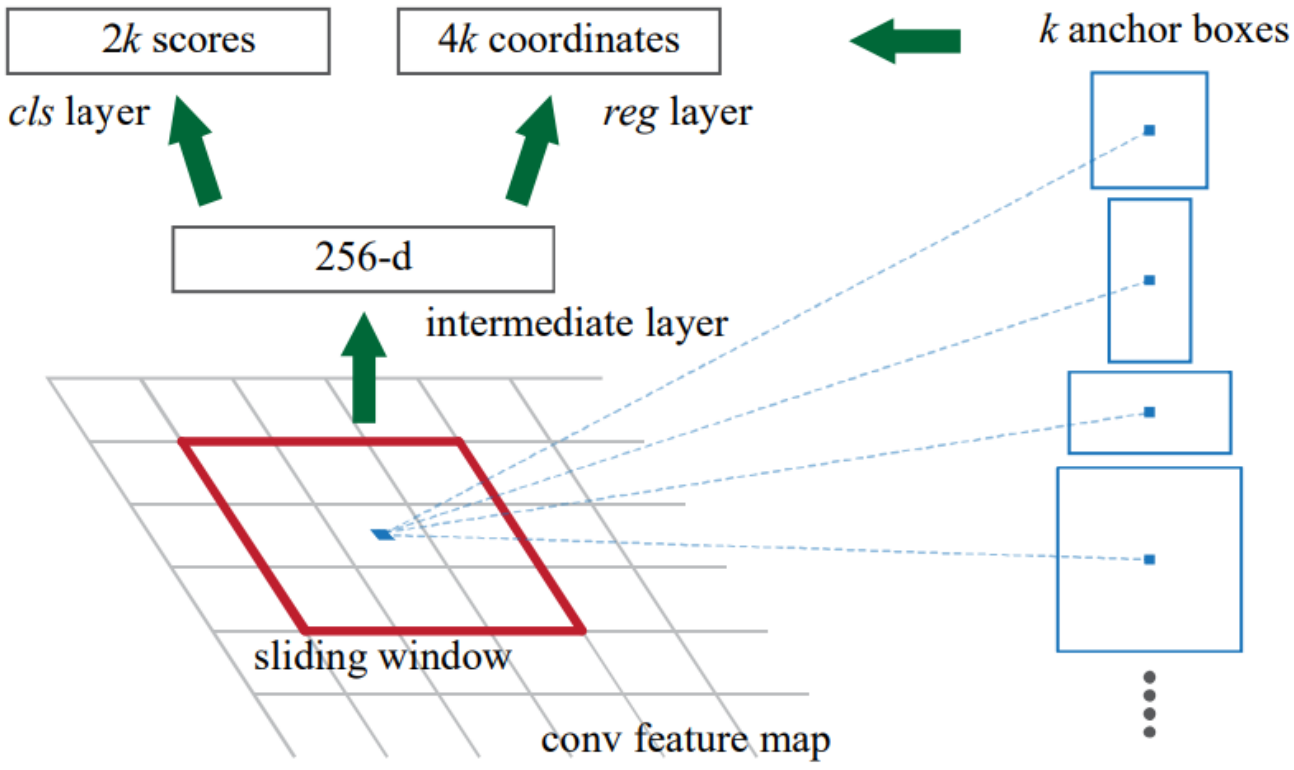

Рисунок 3.2 – Схема роботи RPN

Кожне ковзне вікно накладається на вектор нижньої вимірності. Цей вектор подається на два повністю пов’язані шари — шар коробкової регресії (reg) і шар коробкової класифікації (cls). У даній роботі ми використовуємо n = 3, зауважуючи, що ефективне рецептивне поле на вхідному зображенні велике (171 і 228 пікселів для ZF і VGG відповідно). Слід зазначити, що оскільки мережа

функціонує у вигляді розсувного вікна, повністю підключені шари використовуються спільно в усіх просторових місцях. Ця архітектура реалізована за допомогою рівня n × n, за яким слідують два однотипні рівні 1 × 1 (для reg і cls відповідно). ReLU застосовуються до виходу загорткового шару n × n. Приклади виявлення об’єктів за допомогою методи пропозицій RPN наведено на рисунку 3.3. Метод мережа регіональних пропозицій дозволяє виявляти об’єкти в широкому діапазоні масштабів і пропорцій.

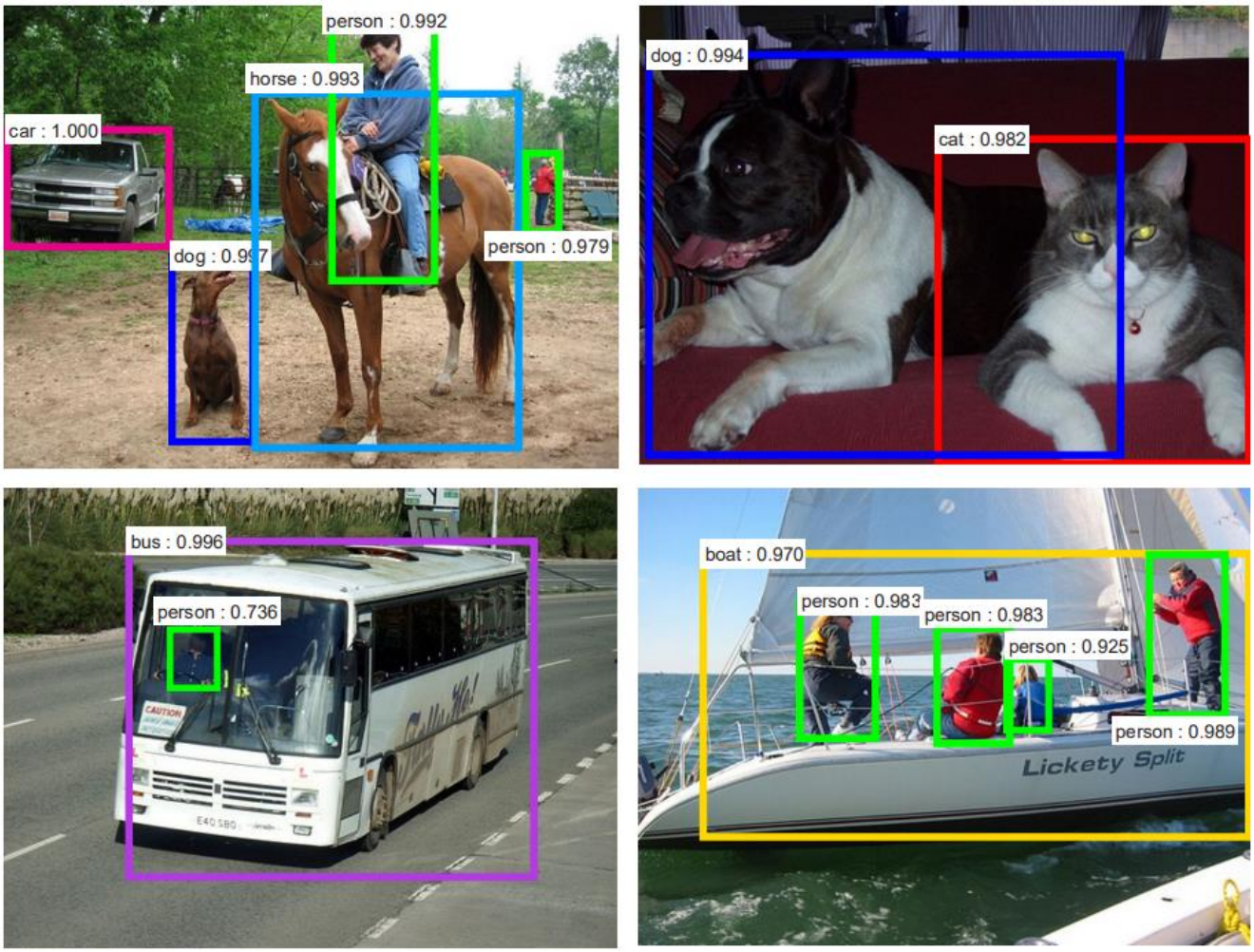

Рисунок 3.3 – Приклади результатів роботи RPN

Модель Fast R-CNN спільно з RPN дозволяє застосувати функції згорткових мереж для пропозиції регіону та виявлення об’єктів. І RPN, і R-CNN, навчаються незалежно і змінюють свої рівні конверсії різними способами. Тому

щоб вони запрацювали потрібна техніка, яка дозволить ділитися рівнями конверсій між двома мережами, а не вивчати дві окремі мережі. Це не так просто, як визначити одну мережу, яка включає як RPN, так і Fast R-CNN, а потім оптимізувати її разом із зворотним поширенням. Причина полягає в тому, що навчання Fast R-CNN залежить від пропозицій фіксованого об'єкта. Саме тому у обраному алгоритмі генерації підписів до зображення використовується прагматичний чотирьох етапний алгоритм навчання для вивчення спільних функцій за допомогою поперемінної оптимізації. Спочатку навчається RPN, як зазначено у [14]. Ця мережа ініціалізується попередньо підготовленою моделлю ImageNet і повністю налаштована для завдання пропозиції регіону. Далі на другому кроці ми навчаємо окрему мережу виявлення за допомогою Fast R-CNN, використовуючи пропозиції, згенеровані у попередньому кроку RPN. Ця мережа виявлення також ініціалізується попередньо навченою моделлю ImageNet, та на даної стадії навчання обидві мережі не мають спільні згорткові рівнів. На третьому кроці навчання використовується мережа детекторів для ініціалізації навчання RPN, але спільні шари конверсії виправляються та точно налаштовуються шари, унікальні для RPN. Після чого обидві мережі вже спільно використовують згорткові рівні. Нарешті, залишаючи спільні згорткові шари фіксованими, стає можливим тонке налаштування Fast R-CNN шарів. Таким чином, обидві мережі мають однакові рівні конверсії та утворюють єдину мережу.

3.3 Огляд алгоритму трансформатора об'єктних відносин

Модель Transformer складається з кодера та декодера, обидва з яких складаються із стеку шарів (в нашому випадку 6 шарів). Для підписів зображень архітектура методу використовує вектори ознак із детектора об’єктів як вхідні дані і генерує послідовність слів (тобто підпис зображення) як вихідні дані. Кожен вектор ознак зображення спочатку обробляється через вхідний шар вбудовування, який складається з повністю підключеного шару для зменшення розміру з 2048 до dmodel = 512, а потім ReLU і випадаючий шар. Кожен шар

кодера складається з шару самоуваги з кількома головками, за яким слідує невелика нейронна мережа з прямим зв'язком.

Механізми масштабованої точкової уваги (Scaled Dot-Product Attention) та багатоголової уваги (Multi-Head Attention), що застосовуються в моделі Transformer складаються з кількох рівнів уваги, що працюють паралельно. Схему роботи цих механізмів уваги зображено на рисунку 3.4.

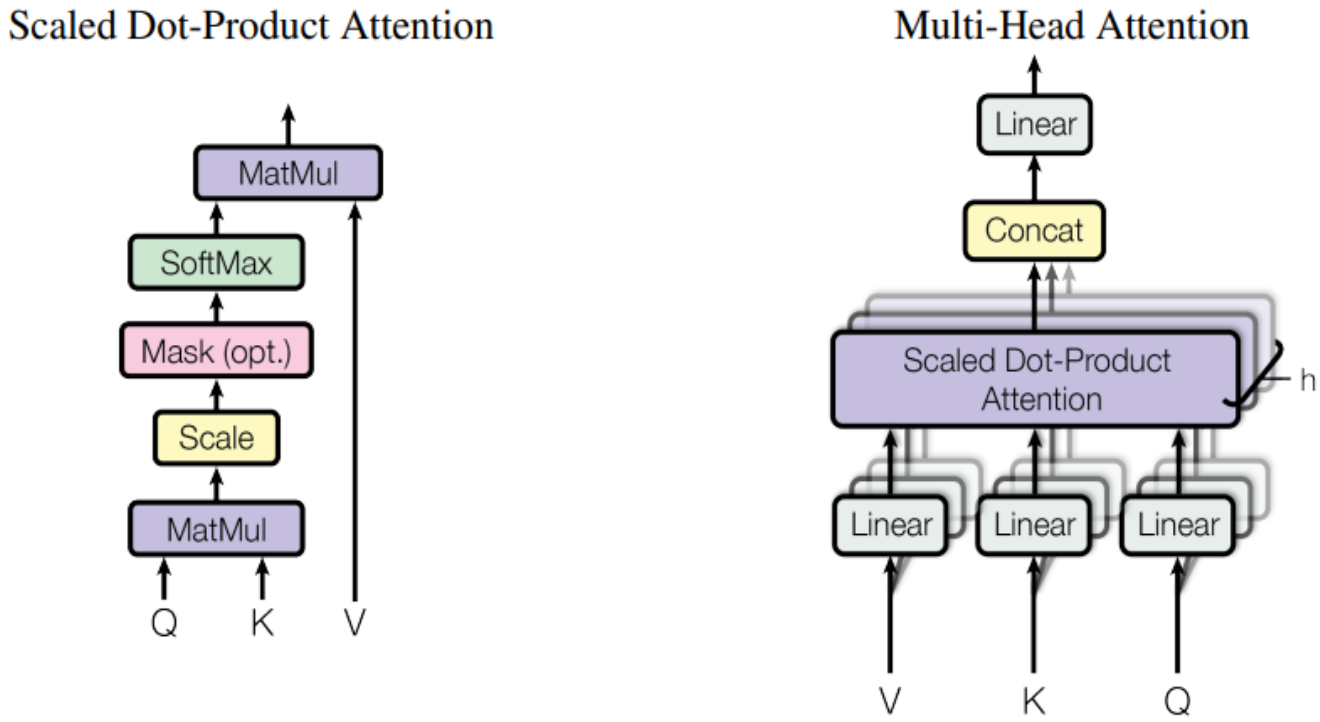

Рисунок 3.4 – Огляд архітектури Object Relation Transformer

Сам шар самоуваги складається з 8 однакових головок. Кожна голова уваги спочатку обчислює запити Q, ключі K і значення V для N токенів наступним чином:

$$Q = XW_Q, K = XW_K, V = XW_V, \tag{1}$$

де X містить усі вхідні вектори x1...xN, зібрані в матрицю;

W_Q, W_K і W_V – це вивчені проекційні матриці.

Потім обчислюються вагові коефіцієнти уваги для особливостей зовнішнього вигляду:

$$\Omega_A = \frac{QK^T}{\sqrt{d_k}} \tag{2}$$

де Ω_{A} – матриця ваги уваги N × N, елементи якої ω_{A}^{mn} коефіцієнти уваги між m-им і n-им маркерами. У якості постійного коефіцієнту масштабування d_k = 64, який є розмірністю векторів ключа, запиту та значення.

Тоді вихід голови розраховується як:

$$\text{head}(X) = \text{self-attention}(Q, K, V) = \text{softmax}(\Omega_A)V \tag{3}$$

Рівняння з 1 по 3 розраховуються для кожної голови окремо. Потім вихідні дані всіх 8 голов об’єднуються в один вихідний вектор і помножуються на вивчену проекційну матрицю W_0:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{4}$$

Наступним компонентом рівня кодера є точкова мережа прямої подачі (FFN), яка застосовується до кожного виходу рівня уваги.

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

де W_1, b_1 і W_2, b_2 – ваги та зміщення двох повністю з'єднаних шарів. Крім того, до виходів самоуважності та шару прямої подачі застосовуються з'єднання пропуску та шару норми (layer-norm).

Потім декодер використовує згенеровані маркери з останнього шару кодера як вхідні дані для створення тексту підпису. Оскільки розміри вихідних маркерів кодера Transformer ідентичні токенам, які використовуються в оригінальній реалізації Transformer, ми не вносимо жодних змін на стороні декодера[7].

У обраній моделі також включається відносна геометрія, змінюючи матрицю ваги уваги Ω_A в рівнянні 2. Ми множимо вагові коефіцієнти уваги ω_A^{mn}двох об'єктів m і n на вивчену функцію їх відносного розташування та розміру.

Спочатку ми обчислюємо вектор переміщення λ(m, n) для обмежуючих прямокутників m і n на основі їх геометричних об'єктів (x_m, y_m, ω_m, h_m) і (x_n, y_n, ω_n, h_n) (координати центру, ширини та висоти), як

$$\lambda(m,n) = \left(\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|y_m - y_n|}{h_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right) \quad (6)$$

Геометричні ваги уваги потім обчислюються, як

$$\omega_G^{mn} = ReLU\left(\mathrm{Emb}(\lambda)W_G\right) \quad (7)$$

де $Emb(\lambda)$ обчислює високорозмірний коефіцент вбудовування слідуючи функціям PE_{pos}, де синусоїдальні функції рахуються для кожного значення $\lambda(m,n)$.

Щоб спроектувати до скаляру і застосувати нелінійність ReLU ми помножаємо коефіцент вбудовування на вектор W_G. Геометричні ваги уваги W_G^{mn} потім включаються в механізм уваги відповідно до

$$\omega^{mn} = \frac{\omega_G^{mn}\exp(\omega_A^{mn})}{\sum_{l=1}^{N}\omega_G^{ml}\exp(\omega_A^{ml})} \quad (8)$$

де W_A^{mn} – це вагові показники уваги на основі зовнішнього вигляду з рівняння 2; ω^{mn} – є новими комбінованими вагами уваги.

На рисунку 3.1 показаний огляд запропонованого алгоритму підписання зображень. Спочатку використовується детектор об'єктів, щоб отримати елементи зовнішнього вигляду та геометрії з усіх виявлених об'єктів на зображенні. Після цього Object Relation Transformer використовується для створення підпису до зображення. Діаграма реляційного кодування обмежувальної рамки на малюнку показує шар самоуваги з кількома головами трансформатора об'єктних відносин.

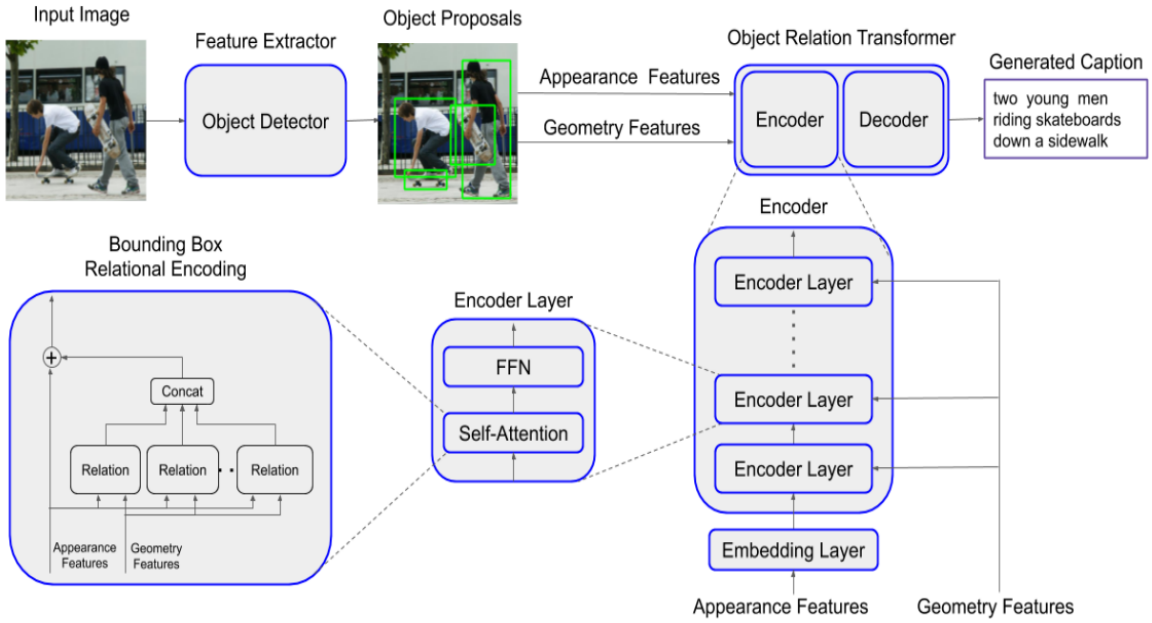

Рисунок 3.5 – Архітектура моделі Object Relation Transformer

Вихід голови можна розрахувати як

$$\text{head}(X) = \text{self-attention}(Q, K, V) = \Omega V \tag{9}$$

де Ω_A – матриця ваги уваги N × N, елементи якої взяті з матриці ω_A^{mn}. Рівняння з 6 по 9 представлені за допомогою коробки відношень.

4 ВИБІР НАБОРУ ДАНИХ ДЛЯ НАВЧАННЯ МОДЕЛІ

4.1 Огляд набору даних Microsoft COCO 2014

Набір даних MS COCO — це широкомасштабний набір даних виявлення, сегментації та субтитрів, опублікований Microsoft. Інженери машинного навчання та комп'ютерного зору широко використовують набір даних COCO для різних проектів комп'ютерного зору.

Набір даних COCO складається з 123 293 зображень, розділених на навчальний набір із 82 783 зображень і набір для перевірки 40 504 зображень. Додатково 40 775 зображень проходять для тестування. Зображення анотовані п'ятьма створеними людьми підписами (дані C5), хоча 5000 випадково вибраних тестових зображень мають по 40 підписів (дані C40).

Людські оцінки COCO були зібрані за допомогою Amazon Mechanical Turk (AMT) з метою оцінки заявок на конкурс COCO Captioning Challenge у 2015 року. Загалом було зібрано 255 000 людських суджень, які представляли собою три незалежні відповіді на п'ять різних питань, які були поставлені стосовно 15 конкурсних робіт, а також людські та випадкові роботи (всього 17). Запитання охоплюють параметри загальної якості субтитрів (M1 - M2), правильності (M3), детальності (M4) і помітності (M5), як детально описано в таблиці 1. Для парного рейтингу (M1, M2 і M5) кожен запис була оцінена з використанням тієї ж підмножини з 1000 зображень із тестового набору C40. Усі оцінювачі AMT складалися з носіїв англійської мови, які проживають у США, занесених до білого списку з попередньої роботи. Метричні бали для конкурсних робіт були отримані від організаторів COCO, використовуючи для обчислення метрику SPICE. Методологію SPICE було виправлено перед оцінкою на COCO.

4.2 Огляд наборів даних Flickr

Набір даних Flickr8K містить 8092 зображення, анотовані п'ятьма створеними людьми посиланнями на кожному. Зображення були підібрані вручну, щоб зосередитися в основному на людей і тварин, які виконують дії.

Набір даних також містить оцінені показники якості людини для 5822 підписів із оцінками від 1 («вибраний підпис не пов’язаний із зображенням») до 4 («вибраний підпис описує зображення без жодних помилок»). Кожен заголовок оцінювався трьома експертами-оцінювачами, отриманими від групи носіїв мови. Усі оцінені підписи були отримані з набору даних, але асоціація із зображеннями була виконана за допомогою системи пошуку зображень. При навчанні Image Captioning моделей виключається 158 правильних пар зображення-підписи, де кандидатський підпис з’являється в наборі посилань, що зменшує всі показники кореляції, але не впливає непропорційно на жодну метрику.

Набір даних Flickr30K є наступною більш покращеною версією Flickr8K, що став одним із стандартних еталонів для опису зображень на основі речення. Цей набір даних містить 244 тис. ланцюжків кореферентів і 276 тис. обмежувальних прямокутників із вручну для кожного з 31 783 зображень і 158 915 підписів англійською мовою (п’ять на зображення) у вихідному наборі даних. Фрази, які належать до одного ланцюга кореферентів, мають однаковий ідентифікатор ланцюжка. Кожна фраза пов’язана з одним або кількома типами, які відповідають грубим категоріям, описаним у нашій статті. Фрази типу «notvisual» мають нульовий ідентифікатор ланцюга і повинні вважатися набором однотонних ланцюжків кореферентності, оскільки ці фрази не були анотовані.

4.3 Огляд наборів даних Composite Dataset

Існує додатковий набір даних із 11 985 людських суджень щодо підписів Flickr 8K, Flickr 30K та COCO, який називають зведеним набором даних (Composite Dataset). У цьому наборі даних підписи були оцінені за допомогою AMT за шкалою правильності від 1 («Опис не має відношення до зображення») до 5 («Опис ідеально пов’язаний із зображенням»). Субтитри-кандидати були отримані з людських довідкових підписів і двох моделей субтитрів [15].

4.4 Огляд наборів даних PASCAL-50S

Щоб створити набір даних PASCAL-50S, 1000 зображень із набору речень UIUC PASCAL, які спочатку містили п’ять підписів на одне зображення, були анотовані 50 підписами кожне за допомогою AMT. Вибрані зображення представляють 20 класів, включаючи людей, тварин, транспортні засоби та предмети побуту.

Набір даних також включає людські судження про понад 4000 кандидатних пар речень. Однак, на відміну від попередніх досліджень, працівників AMT не просили оцінювати підписи щодо зображень. Фрази, які належать до одного ланцюга кореферентів, мають однаковий ідентифікатор ланцюжка. Кожна фраза пов’язана з одним або кількома типами, які відповідають грубим категоріям, описаним у нашій статті. Фрази типу "notvisual" мають нульовий ідентифікатор ланцюга "0" і повинні вважатися набором однотонних ланцюжків кореферентності, оскільки ці фрази не були анотовані. Якщо еталонні підписи відрізняються за якістю, цей підхід може внести більше шуму в процес оцінки, однак відмінності між цим підходом і попередніми підходами до оцінки людини не вивчені. Для кожної кандидатської пари речень (B,C) були зібрані оцінки щодо 48 із 50 можливих опорних підписів. Пари пропозицій-кандидатів були створені як із підписів людини, так і з моделі, пари чотирма способами: людина-правильна (HC), людина-неправильна (HI), людина-модель (HM) та модель-модель (MM).

4.5 Висновки до розділу

Після огляду найпопулярніших і сучасних наборів даних для навчання моделі трансформатора об'єктних відносин було обрано Microsoft COCO датасет. Цей набір даних є безплатним для завантаження, він містить достатню кількість потрібних для навчання даних (понад 330 тисяч зображень та білш ніж 200 тисяч підписів до них). З набором даних COCO Microsoft представила візуальний набір даних, який містить величезну кількість фотографій, що зображують звичайні

об’єкти в складних повсякденних сценах. Це відрізняє COCO від інших наборів даних розпізнавання об’єктів, які можуть бути специфічними секторами штучного інтелекту. Такі сектори включають класифікацію зображень, локалізацію рамки об’єкта або семантичну сегментацію на рівні пікселів. Тим часом анотації COCO в основному зосереджені на сегментації кількох окремих екземплярів об’єктів. Цей більш широкий фокус дозволяє використовувати COCO в більшій кількості випадків, ніж інші популярні набори даних, такі як CIFAR-10 і CIFAR-100, Flickr та PASCAL-50S.

5 ДОСЛІДЖЕННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

5.1 Реалізація та навчання моделі

Після аналізу існуючої наукової літератури обраний Image Captioning Object Relation Transformer алгоритм було реалізовано на мові програмування Python та фреймворку PyTorch в середовищі проектування PyCharm. Алгоритм використовує наступні бібліотеки: h5py, scikit-image, typing, pyemd, gensim. Для оцінювання згенерованих написів були використанні бібліотеки cider і coco-caption, які дозволяють отримувати оцінки згенерованих речень в різних метриках включаючи метрики CIDER, BLEU, CPICE, ROUGE, які є де факто стаднартними метриками для оцінювання Image Captioning алгоритмів. Навчання моделі проводилося на датасеті MSCOCO. Цей датасет містить 113 000 навчальних зображень із 5 анотованими підписами для кожного зображення, і є одним із стандартних датасетів для оцінювання Image Captioning моделей.

Під час навчання моделі треба спочатку завантажити список попередньо оброблених підписів COCO, який наданий у форматі JSON в файлі dataset_coco.json. Цей файл містить попередньо оброблені субтитри, а також стандартні розділення «train-val-test». Далі ми обробляємо даний файл з метою найти усі слова, які зустрічаються менше 6 разів, зі спеціальним маркером UNK і створюємо словниковий запас для всіх слів, що залишилися. Інформація про зображення та словник і про дані дискретних підписів записуються в два окремих файли. Це дозволить попередньо обробити набір даних і отримати кеш для розрахунку оцінки CIDER. Під час навчання моделі на додаток до перехресної втрати ентропії також проводиться оцінка отриманих результатів за показниками BLEU, METEOR та CIDEr. У кінці навчання було отримано оцінку 115 за метрикою CIDEr.

Після тренування з використанням перехресної втрати ентропії додаткове самокритичне навчання дає значний приріст оцінки CIDEr-D. Приріст у оцінках метрик після експерименту з участю п’яти тисяч тестових зображень для двох моделей трансформатора об'єктних відносин, де перша була навчена без

використання самокритичного навчання, а друга є та сама модель (створена копія) але вже після самокритичного навчання, наведено у таблиці 5.1.

Таблиця 5.1 – Порівняння оцінок результатів роботи алгоритму без і з самокритичним навчанням

Назва метрики	Оцінки без самокритичним навчанням	Оцінка з самокритичним навчанням
CIDEr-D	115.37	128.3
SPICE	21.24	22.6
BLEU-1	76.63	80.5
BLEU-4	35.49	38.6
METEOR	27.98	28.7
ROUGE-L	56.58	58.6

Але навіть після закінчення навчального процесу моделі було виявлено помилки в її роботі. Деякі приклади незвичайних об’єктів, які не були правильно ідентифіковані, включають: паркомат, манекен одягу, капелюх-парасольку, трактор та малярську стрічку. Ця проблема також помітна, хоч і в меншій мірі, у рідкісних відносинах та атрибутах. Цікаво що спостереження полягало в тому, що створені підписи, як правило, менш описові та менш дискурсивні, ніж основні підписи правди. Наведені вище результати та спостереження можуть бути використані для визначення пріоритетів майбутні зусилля щодо підписання зображень.

5.2 Результати роботи алгоритму і порівняння отриманих результатів

Приклади надписів для зображень згенерованих навченою моделлю можна побачити на рисунках 4.1 та 4.2.

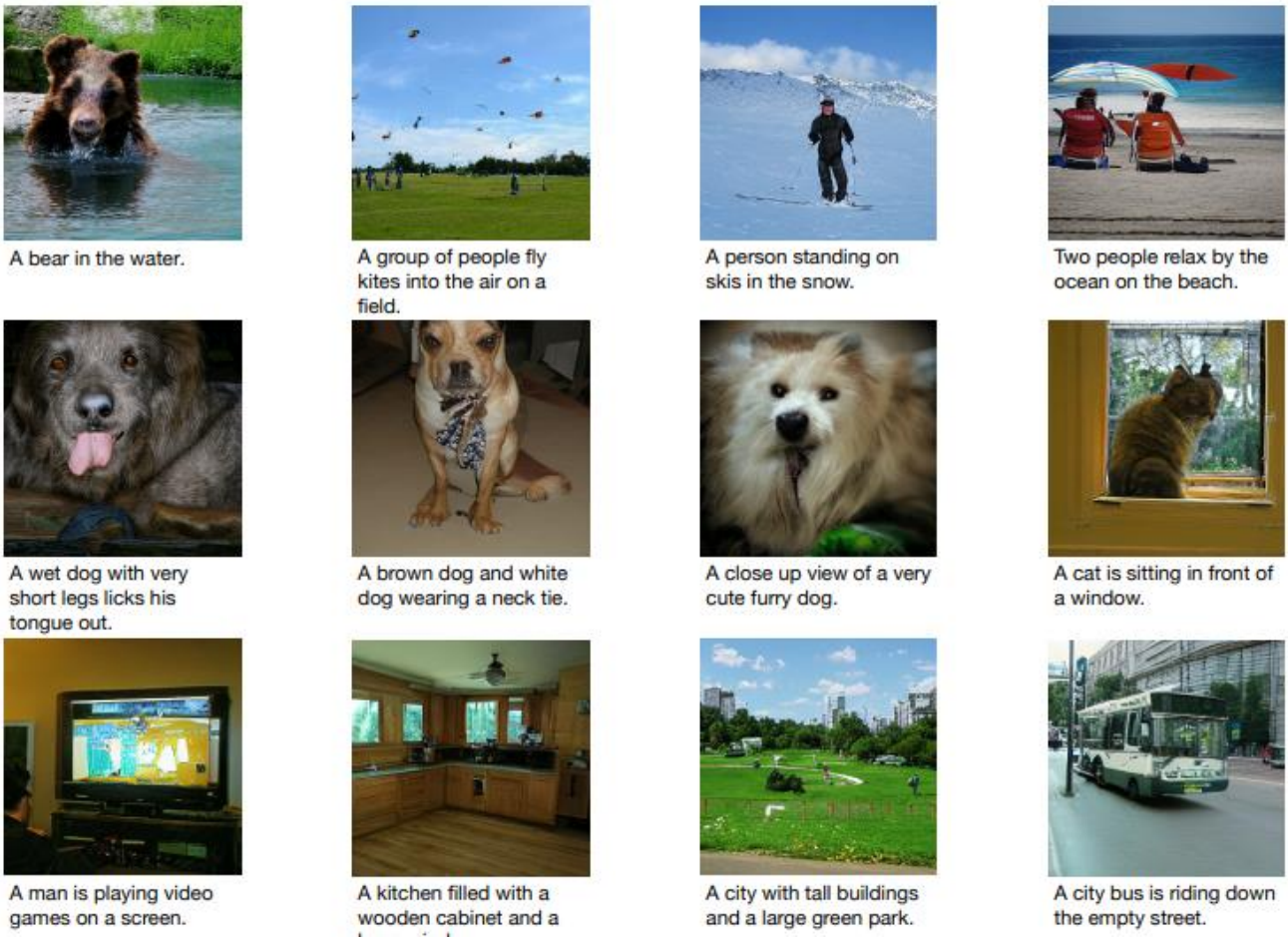

Рисунок 4.1 – Підписи до зображень згенеровані алгоритмом Object Relation Transformer

Отримана модель дуже схоже на модель Standard Transformer але різниця полягає саме у механізмі геометричної уваги [19], який можна розглядати як заміну позиційного кодування в схожій мережі. Для розміру коробки ми просто обчислюємо площу кожного обмежуючого прямокутника та впорядковуємо від найбільшого до найменшого.

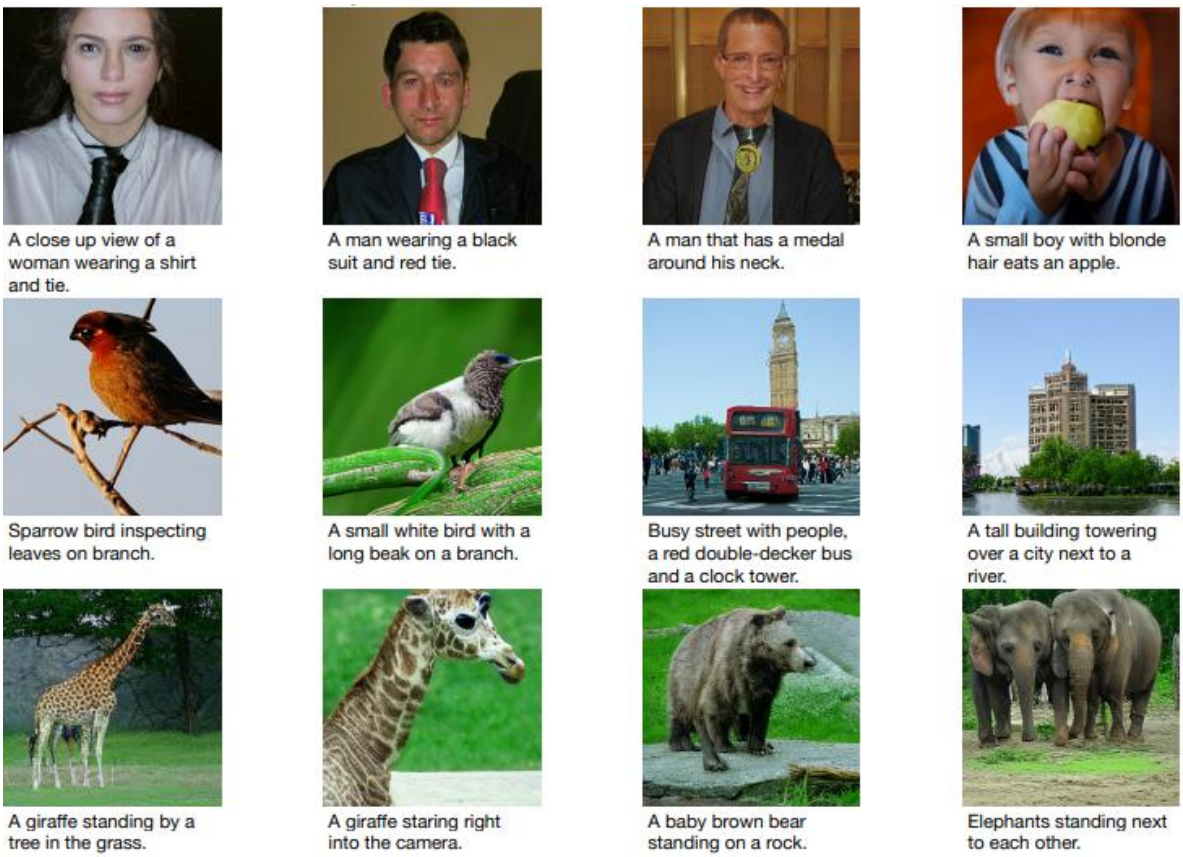

Рисунок 4.2 – Підписи до зображень згенеровані алгоритмом Object Relation Transformer

Для порівняння з результатів роботи навченої моделі було обрано наступні моделі: Virtex [13], From Captions to Visual Concepts and Back [14], ClipCap-Transformer [15], ClipCap(MLP + GPT2 tuning) [15], RDN(Reflecting Decoding Network) [16], RefineCap [17], CNN + LSTM, Standard Transformer. Результати порівняння можно побачити в таблиці 4.2

Таблиця 4.2 – Порівняння оцінок алгоритмів Image Captioning з алгоритмом Object Relation Transformer

Назва алгоритму	BLEU-4	CIDER	METEOR	SPICE
1	2	3	4	5
From Captions to Visual Concepts and Back	25.7	–	23.6	–

Продовження таблиці 4.2

1	2	3	4	5
Virtex	–	94	–	18.5
ClipCap-Transformer	33.53	113.08	27.45	21.05
ClipCap (MLP)	32.15	108.35	27.1	20.12
RefineCap	37.8	127.2	28.3	22.5
RDN	37.3	125.2	28.1	–
LSTM	32.9	106.6	26.5	19.9
Standard Transformer	32.8	111	27.5	20.9
Object Relation Transformer	38.6	128.3	28.7	22.6
Transformer_NSC	39.4	129.6	28.9	22.8
Meshed-Memory Transformer	39.1	131.2	29.2	22.6

Треба зазначити що не всі моделі мають оцінку в обраних метриках. Для таких моделей в ячейках конкретних метриках використовується прочерк. Порівняння отриманих оцінок для обраної моделі порівняно було порівняно з оцінками інших Image Captioning моделей навчених на тому ж самому MSCOCO дата сеті. Хоча експериментально було показано, що BLEU і ROUGE мають нижчу кореляцію з людськими судженнями, ніж інші показники, загальною практикою в літературі щодо підписів зображень є повідомлення про всі вищезгадані показники.

З наведеної таблиці можна побачити результати роботи геометричного шару уваги, якщо порівняти обрану модель з оцінками моделі схожої моделі Standard Transformer. Також можна помітити різницю між LSTM моделлю і Object Relation Transformer. В обох цих моделях як частина шифрувальника використається конволюційні нейроні мережі, тобто основною різницею цих

обох моделей є саме частина декодування. З отриманих різниць в метриках можна побачити, що дописи згенеровані за допомогою моделі Transformer є повноцінні як з точки зори повноти так і з точки зору влучності.

Слід зазначити, що порівняння складається з набору показників оцінки, обчислених для кожної моделі на основі кожного зображення, а також агрегованих для всіх зображень. Він включає результати парних t-тестів, які перевіряють статистично значущі відмінності між оціночними показниками, отриманими в результаті кожної з моделей.

Моделі LSTM, Transformer та Object Relation Transformer використовують одну й ту саму CNN для в частині кодування. Але лише Object Relation Transformer включає механізм геометричну увагу. Внесок трансформатора об’єктних відносин невеликий для METEOR, але значний для показників CIDer-D і BLEU. Загалом ми бачимо найбільше покращень саме для метрик CIDer-D і BLEU-4. Але слід зазначити що отримані оцінки в ході дослідження для обраного датасету є несуттєво нижчими за дві існуючих моделі Transformer_NSC та Meshed-Memory Transformer, але як буде зазначено далі то у моделі трансформатора об'єктних відносин геометричну увагу не включається в шари перехресної уваги декодера між об’єктами та словами, що є існуючим потенціалом для подальшого росту оцінок отриманих обраними метриками і дає основи вважати, що у потенціалі модель Object Relation Transformer при подальшому покращені дозволить отримати більш високі оцінки метрик ніж моделі моделі Transformer_NSC та Meshed-Memory Transformer.

Об’єктом цього експериментального дослідження є модель генерації підписів до зображення Object Relation Transformer, яка є модифікацію звичайного Transformer, спеціально адаптовану для завдання підпису зображень. Обрана модель трансформатора кодує двовимірні відносини розташування та розміру між виявленими об’єктами на зображеннях, ґрунтуючись на підході підписання зображень «знизу вгору» та «зверху вниз». Отримані результати на наборі даних MS-COCO демонструють, що модель трансформатора об'єктних

відносин дійсно отримує вигоду від включення інформації про просторові відносини, що найбільш очевидно, якщо порівнювати отримані оцінки метрик і порівняти їх з іншими моделями без цієї властивості. Результати роботи навченої моделі також представили якісні приклади того, як включення цієї інформації може дати результати підписів, що демонструють кращу просторову обізнаність. Наразі обрана модель враховує лише геометричну інформацію на етапі кодера. Напрямом для наступних досліджень може бути модель, яка включить геометричну увагу в шари перехресної уваги декодера між об’єктами та словами, що може бути реалізовано шляхом явного зв’язування декодованих слів із обмежуючими рамками об’єкта. Це повинно призвести до додаткового підвищення продуктивності, а також до покращення інтерпретації моделі.

ВИСНОВКИ

В результаті роботи було проведено дослідження методів Image Captioning. Була проаналізована відповідна наукова література. Були проаналізовано основні типи алгоритмів і їх сильні і слабкі сторони використання.

Обраний алгоритм трансформера об'єктних відносин(Object Relation Transformer), структуру якого було описано у роботі дозволяє отримати state-of-the-art продуктивність на багатьох датасетах, таких як COCO Captions, flickr8k, flickr32k та інших. Обраний трансформер кодує 2D позицію і співвідношення розмірів між виявленими об’єктами на зображеннях, спираючись на підхід до підписів зображень знизу вгору та вниз.

Наразі вивчена модель враховує лише геометричну інформацію на етапі кодера. Одним з наступних кроків можливо включити геометричну увагу в шари перехресної уваги декодера між об’єктами та словами.

При подальшому вивчені цієї теми можливо додавання нових джерел інформації та розширення існуючого функціоналу шляхом використання інших алгоритмів і нейронних мереж в фрейморку енкодеру-декодеру.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. E. Erdem, M. Kuyu, S. Yagcioglu, A. Frank, L. Parcalabescu, B. Plank, A. Babii, O. Turuta, A. Erdem, I. Calixto, E. Lloret, E.-S. Apostol, C.-O. Truica, B. Sandrih, A. Gatt, S. Martincic-Ipsic, G. Berend, and G. Korvel. 2022. Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning. Journal of Artificial Intelligence Research, in press.

2. MF Bondarenko, ZV Dudar, IA Ephimova, VA Leshchinsky, S Yu Shabanov-Kushnarenko. 2004. About brain-like computers.

3. Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In: European conference on computer vision, Berlin.

4. Hoiem, D., Divvala, S., Hays, J.: Pascal voc 2009 challenge. In: PASCAL challengeworkshop in ECCV (2009)

5. А.Л. Ерохин, А.С. Нечипоренко, А.С. Бабий, А.П. Турута. 2016. Применение глубоких сверточных нейронных сетей для классификаци риноманометрических данных.

6. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp. 2048-2057.

7. Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

8. Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization

9. Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In CVPR.

10. A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3128–3137, 2015.

11. Peter Anderson, Basura Fernando, Mark Johnson and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In ECCV.

12. J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4565–4574, 2016.

13. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

14. Levy, R.: Expectation-based syntactic comprehension. Cognition 106(3) (2008).

15. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3156–3164, 2015.

16. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057, 2015.

17. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015.

18. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. (2015).

19. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

20. Karan Desai, Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations, 2020.

21. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, Geoffrey Zweig. From Captions to Visual Concepts and Back, 2015.

22. Yekun Chai, Shuo Jin, Junliang Xing. RefineCap: Concept-Aware Refinement for Image Captioning, 2021.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

1. E. Erdem, M. Kuyu, S. Yagcioglu, A. Frank, L. Parcalabescu, B. Plank, A. Babii, O. Turuta, A. Erdem, I. Calixto, E. Lloret, E.-S. Apostol, C.-O. Truica, B. Sandrih, A. Gatt, S. Martincic-Ipsic, G. Berend, and G. Korvel. 2022. Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning. Journal of Artificial Intelligence Research, in press.

2. MF Bondarenko, ZV Dudar, IA Ephimova, VA Leshchinsky, S Yu Shabanov-Kushnarenko. 2004. About brain-like computers.

5. А.Л. Ерохин, А.С. Нечипоренко, А.С. Бабий, А.П. Турута. 2016. Применение глубоких сверточных нейронных сетей для классификаци риноманометрических данных.