Швець Сергій Валерійович, магістр

Приватний вищий навчальний заклад "Європейський університет"

0009-0006-0060-7760

**ЕФЕКТИВНІСТЬ ТА ОСОБЛИВОСТІ СУЧАСНИХ МОДЕЛЕЙ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ**

**Проблема класифікації зображень**

У сучасному світі бурхливий розвиток електронної комерції ставить нові виклики перед когнітивними та обчислювальними можливостями інформаційних систем. Постійно зростаючий об'єм інформації та зростаючі вимоги до валідності та організації даних вимагають автоматизації процесів перевірки та класифікацій, у тому числі і щодо графічних зображень. Такі задачі вимагають когнітивних спроможностей від алгоритмів, що неможливо без використання сучасних алгоритмів побудованих на основі глибинного навчання.

**Фундамент глибинного навчання**

Алгоритми глибинного навчання беруть натхнення із біологічних принципів побудови мозкового апарату, а саме будова нейрона та спосіб організації нейронів у мережу. Ідея перцептрона, штучного нейрона, була запропонована Френком Розенблатом у 1957 році. Запропонована ним модель добре справляється із лінійно-роздільними даними. Для більш складних задач модель перцептрона відповідно не застосовна, і у свою чергу було запропоновано багатошарові нейронні мережі. Це у свою чергу створило фундамент для сучасного глибинного навчання, що здатне вирішувати у тому числі і задачі комп'ютерного зору та класифікації зображень.

Наразі запропоновано багато алгоритмів класифікації зображень, і постає питання їх еталонного співставлення для об'єктивної оцінки їх ступеня ефективності. Як рішення, дослідниками із Прінстонського університету, було запропоновано і наразі підтримується, еталонний набір даних ImageNet. Станом на зараз він нараховує 14 мільйонів зображень у 20 тис. категоріях. Щороку відбувається змагання ILSVRC (ImageNet Large Scale Visual Recognition Challenge) - де дослідники з усього світу змагаються в точності класифікації зображень, використовуючи еталоні дані ImageNet. Це змагання дало надзвичайний поштовх бурхливому розвитку алгоритмам розпізнавання та класифікації зображень.

**Методологія оцінювання ефективності ILSVRC**

Для оцінювання ефективності у змаганні ILSVRC використовують декілька показників

*Топ-1 або Top-1 Accuracy*

Це відсоток випадків, у яких запропонована модель надала класифікацію вірно. Тут треба розуміти вірність класифікації як співпадіння еталонного класу із найбільш імовірним класом згідно передбачення модели що тестується.

*Топ-5 або Top-5 Accuracy*

Більш м'який критерій, що відповідно відповідає відсотку знаходження еталоного класу серед пяти найбільш імовірних із запропонованих моделлю. Це важливий критерій оскільки у випадку складних зображень когнітивно вірних міток може бути декілька.

*Матриця помилок, або Confusion matrix*

Матриця розмірністю NxN (N - кількість класів), що відповідно відображає кількість випадків, коли еталонний клас i було класифіковано моделлю як j клас.

Відповідно, матриця корисна для візуалізації заплутаності моделі та аналізу параметрів True Positive, True Negative, False Positive, False Negative.

*Cross-entropy loss*

Ця метрика розраховується відповідно до формули



Де

* *M* — кількість класів
* N — кількість зразків
* *yo*,*c -* це індикатор (0|1), який показує, чи відповідає зразок класу *c*
* *po*,*c -* відповідно передбачена ймовірність приналежності зразка *o* до класу *с*

### Відповідно ця метрика надає усереднену оцінку ефективності модели загалом.

*Оцінювання*

Для розробки дослідники використовують валідаційний набір даних, щоб виконати налаштування параметрів модели. Тестовий набір використовуються під час остаточного оцінювання запропонованої

**Поточні результати**

Завдяки ImageNet та ILSVRC дослідники зі всього світу отримали можливість дослідити ефективність власних моделей на великорозмірних тестових даних. Як результат, запропоновані моделі класифікації зображень досягають надзвичайно високих ефективностей.



Поточний стан ефективності наявних моделей згідно із ILSVRC

Джерело - <https://paperswithcode.com/sota/image-classification-on-imagenet>

Розглянемо деякі із моделей лідерів.

**Сучасні моделі, застосовні для класифікації зображень**

*OmniVec - 92,4% Top-1 Accuracy*

Запропонована компанією OpenAI, ця фреймворк є багатомодальним, тобто застосовна для обробки різних типів даних (текст, зображення, відео та аудіо тощо). Для кожного типу даних використовується окремий енкодер, але простір векторів та backbone мережа є загальною, що дозволяє моделі використовувати різні модальності для навчання. Навчання по різних модальностях запропоновано проводити послідовно. Відповідно для різних задач використовуються різні голови (heads). У якості енкодеру може бути застосовано трансформер(Vision Transformer, ViT), згорткову мережу або можливо використання безпосередніх даних. Відповідно, інформація, отримана з різних модальностей координується та інтегрується спільним проміжним шаром.

*CoCa - 91,0% Top-1 Accuracy*

Цей підхід було запропоновано Google. Він поєднує в собі кілька парадигм(single-encoder, dual-encoder та encoder-decoder) у одній модели, що натренована із одночасно із застосуванням методу контрасних втрат та втрат під час генерування підписів (captioning / generative loss). Енкодер представлено трансформером або згортковою мережею, а декодер використовує архітектуру трансформера, що на першому етапі не використовує крос-увагу в унімодальних шарах, а надалі використовує крос-увагу до виходів енкодера зображення для навчання мультимодальних репрезентаций зображення-текст.

*Model Soup BASIC-L - 90,98% Top-1 Accuracy*

Ця методика полягає у тому, щоб об'єднати результати передбачень різних моделей через певні вагові коефіцієнти, усереднення або вагове голосування. Зазвичай використовують моделі, що добре себе зарекомендували у необхідних межах застосувань. Відповідно, вартість обчислень із використанням даного підходу може значним чином перевищувати інші методи, де використовується одна модель.

Проте цей підхід дає високу стійкість рішення і витрати можуть бути виправдані у критичних застосуваннях що вимагають високого ступеня надійності.

**Шляхи покращення**

Не зважаючи на вузкість задачі по класифікації зображень безсумнівними лідерами у рішенні задач класифікації стають фреймворки та моделі що використовують мультимодальний підхід до навчання мереж. Проте, це призводить до збільшення кількості параметрів, зростанню обчислювальної складності та вимагає інноваційного підходу до енкодерів. Отже, можна зробити припущення, що майбутнє покращення ефективності можливо досягти шляхом пошуку більш продуктивних методів для вищесказаних елементів а також шляхом оркестрації існуючих архітектур. Це заслуговує на додаткові дослідження да експерименти.

Література

1. Russakovsky O., Deng J, Su H, Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A. C., Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge URL:<https://arxiv.org/pdf/1409.0575> (дата звернення: 07.05.2024)
2. Srivastava S., Sharma G. OmniVec: Learning robust representations with cross modal sharing URL:<https://arxiv.org/pdf/2311.05709v1> (дата звернення: 07.05.2024)
3. Yu J., Zirui Wang Z. CoCa: Contrastive Captioners are Image-Text Foundation Models URL:<https://arxiv.org/pdf/2205.01917v2> (дата звернення: 07.05.2024)
4. Wortsman M., Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time URL:<https://arxiv.org/pdf/2203.05482v3> (дата звернення: 07.05.2024)