УДК 004.855.5 http://doi.org/10.15407/pp2024.02-03.247

Р.В. Терентьєв, П.А. Іваненко

ПЕРЕДАВАЛЬНЕ НАВЧАННЯ ДЛЯ ПІДВИЩЕННЯ ТОЧНОСТІ КЛАСИФІКАЦІЇ ВІЗУАЛЬНОГО ТРАНСФОРМЕРА НА ОБМЕЖЕНИХ ДАНИХ

У цій статті досліджується ефективність попереднього навчання генеративних моделей на основі візуального трансформера і подальшому навчанню моделі для задач класифікації зображень. Основною проблемою дослідження є погана ефективність навчання візуального трансформера на обмеженій кількості даних. Можна підвищити точність моделі класифікації зображень, використавши передавальне навчання знань, отриманих під час попереднього навчання генеративної моделі на тих самих даних. Для перевірки гіпотези була використана підмножина стандартного набору даних Imagenet, що містить 200 категорій по ~500 зображень. Розмір кожного зображення 64x64 пікселів. Для попереднього навчання генеративної моделі використовуються патчі для маскування сегментів зображення. Процес навчання відновлення замаскованих пікселів зображення змушує модель звертати увагу на контекст навколо видаленої частини, а також на загальні візуальні закономірності. Це приводить до кращого розуміння моделлю візуальної інформації в цілому і допомагає у подальшому навчанні моделі під задачу класифікації. В результаті серії експериментів вдалося досягти покращення точності класифікації зображень з 40% до 44.7%, а також наведено аналіз впливу на нього загального ступеню маскування та розмірності патчів. Додатково в роботі досліджені різні розмірності патчів (2x2, 4x4, 8x8 пікселів) й різний відсоток маскування (20/40/60 відсотків) вхідного зображення та вплив цих параметрів на передавальне навчання.
Ключові слова: візуальні трансформери, генеративні моделі, класифікація зображень, попереднє навчання, передавальне навчання.

R.V. Terentiev, P.A. Ivanenko

TRANSFER LEARNING METHODS FOR INCREASING VISION TRANSFORMER CLASSIFICATION ACCURACY ON SMALL DATASET

This article examines the effectiveness of pre-training generative model based on a visual transformer and subsequent fine tuning for image classification tasks. The main problem of the study is the poor training efficiency of the visual transformer on a limited amount of data. It is possible to improve the accuracy of the image classification model by using transfer learning of the knowledge obtained during the previous training of the generative model on the same data. A subset of the standard Imagenet dataset - Tiny Imagenet was used to test the hypothesis. It contains 200 categories of around 500 images each. The size of each image is 64x64 pixels. For pre-training the generative model, patches are used to mask image segments. The training of restoring masked image pixels forces the model to pay attention to the context around the removed part, as well as to general visual patterns. This leads to a better understanding of visual information by the model as a whole and helps with further fine tuning of the model for the classification task. As a result of a series of experiments, it was possible to achieve an improvement in the accuracy of image classification from 40% to 44.7%, and an analysis of the effect of the overall degree of masking and patch size on it is given. Additionally, impact of different sizes of patches (2x2, 4x4, 8x8 pixels) and different percentages of masking (20/40/60 percent) of the input image were investigated in the paper.
Keywords: vision transformers, generative models, image classification, pre-training, transfer learning.

Вступ

Візуальний трансформер (ViT) [5] – це новаторська архітектура нейронних мереж, яка за останні роки здобула значну популярність у сфері комп'ютерного зору. Моделі з механізмом самоуваги [11] досягли вражаючих результатів у багатьох зада-

чах, таких як обробка природної мови, класифікація зображень, сегментація та детекція об'єктів. Проте навчання таких моделей може бути складним завданням, що потребує ретельного підходу та розуміння специфіки цієї архітектури.

Одна з ключових проблем навчання візуальних трансформерів – це потреба у великих наборах даних. Вона пов'язана з тим, що візуальні трансформери обробляють зображення як послідовність токенів, і для коректного кодування візуальної інформації їм необхідна значна кількість навчальних прикладів. Традиційні згорткові нейронні мережі [6, 9, 10], натомість можуть давати кращі результати з меншими наборами даних завдяки локальному зв'язку: на відміну від повністю пов’язаних нейронних мереж, де кожен нейрон з’єднується з кожним нейроном на попередньому рівні, нейрони згорткової мережі підключаються лише до невеликої локалізованої області вхідних даних.

В цій роботі випробуваний метод попереднього навчання візуального трансформера як генеративної моделі для покращення результатів подальшого навчання моделі для задачі класифікації використовуючи маленьку вибірку даних.

Метод навчання

Для подолання згаданих вище недоліків візуальних трансформерів можна застосувати комбінацію попереднього навчання генеративної моделі й подальшу передачу знань у модель класифікації. Для цього була використана генеративна модель UVCGANv2 [11], яка є комбінацію ViT та U-Net[8], глибокою нейронною мережею, що добре зарекомендувала себе в задачах сегментації та генерування зображень. Попередньо модель була додатково модифікована з метою використання її не лише для генерації, а й для класифікації зображень.

Метод навчання складається з двох етапів. Спочатку ми навчаємо модель на задачі відтворення зображення, яке було попередньо замасковано патчами визначеного розміру. Ступінь маскування є також фіксованою в межах кожного експерименту. Після навчання моделі як генеративної, ця модель продовжує навчання, але вже як класифікатор. Для порівняння результатів була навчена базова модель – модель, яка була одразу навчена як класифікатор, без попереднього навчання генеративної моделі.

Гіпотеза полягає в тому, що під час навчання генеративної моделі, процес відтворення оригінального зображення змушує модель звертати увагу на контекст навколо видаленої частини, а також на загальні візуальні закономірності. Це має приводити до кращого розуміння моделлю візуальної інформації в цілому та покращити результати подальшого навчання моделі як класифікатора. Додатково перевіримо вплив розміру патчів та ступеню маскування на значення функції втрат у відновленні зображення генеративною моделлю. Очікується збільшення функції втрат для більших патчів для всіх ступенів маскування.

Опис моделі

Візьмемо запропоновану модель UVCGANv2 для перетворення зображень. У цій моделі є токен стилю, який в класичній ViT моделі використовується як токен класу. Для модифікації моделі як класифікатора додається один додатковий шар після токену класу. Цей шар використовується для прогнозування ймовірності того, що зображення належить до певного класу.

Оригінальний UVCGANv2 розрахований на зображення розміром 256x256 пікселів, тому конфігурація моделі була зменшена. Замість 4 шарів перетворення вхідного зображення на послідовність візуальних токенів (та навпаки) було використано три шари з розміром каналів 48, 96 та 192 відповідно. Розмір послідовності, яка передається в блоки трансформера – 64 + 1 (токен класу). Кількість блоків трансформера залишилась без змін, а саме 12. Розмір прихованого шару багатошарового перцептрону в кожному блоці трансформера був зменшений до 768. Фінальний розмір моделі – 8M параметрів. Рис. 1 містить спрощену схему моделі.

Також оригінальна модель використовує оптимізацію навчання за допомогою

ReZero [2]. Ця оптимізація не була використана в цій роботі, через те, що початкові експерименти показали погіршення навчання базової моделі під час використання ReZero.

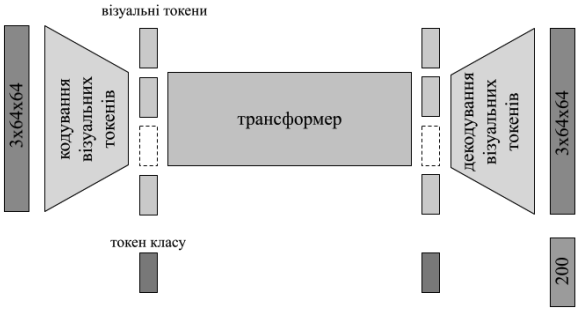

Рис. 1. Спрощена схема модифікованого UVCGANv2 для задачі відновлення зображення та класифікації

Навчання моделі

Дані: Для перевірки гіпотез була використана вибірка даних Tiny Imagenet. Tiny Imagenet – це підмножина вибірки даних ImageNet, що містить 200 категорій замість 1000, та має менший розмір зображень. Ця вибірка даних використовується для дослідницьких задач з комп'ютерного бачення, де потрібна менша кількість даних, але водночас зберігається складність класифікації. Розмір вибірки для тренування – 84000 зображень. Розмір вибірки для валідації – 10000 зображень. Розмір кожного зображення 64x64 пікселів. Додатково під час навчання моделі (як генеративної, так і класифікатора) було використано метод аугментації даних – це процес штучного генерування нових даних на основі вже існуючих. Цей метод використовується для збагачення наборів даних, що, так само веде до кращого узагальнення та більшої стійкості моделей машинного навчання. Завдяки аугментації, модель може "бачити" більше прикладів, що робить її стійкішою до шуму та перепадів у даних. У роботі була використана автоматична аугментація для вибірки даних Imagenet [3].

Функції втрат: Для навчання відтворення зображення була використана функція втрат середня абсолютна похибка між кожним пікселем згенерованого зображення та оригінального зображення. Для навчання класифікації була використана функція втрат перехресна ентропія.

Алгоритм оптимізації: Для мінімізації функції втрат був використаний алгоритм Adam [7]. Він краще працює для тренування моделей з механізмом самоуваги. Adam використовує адаптивне регулювання швидкості навчання для кожного параметра, що дозволяє йому автоматично підлаштовуватися під різні швидкості зближення різних параметрів.

Параметри навчання: Для коректного порівняння кожна з моделей навчалася з аналогічними параметрами. А саме: розмір батчу 128; алгоритм оптимізації моделі – Adam (beta1=0.9, beta2=0.99); навчання впродовж 50 епох; перші 5 епох швидкість навчання збільшується з 1e-9 до 1e-4; для останніх 15 епох швидкість навчання зменшується до 5e-5.

Використання маленької швидкості навчання на початку тренування покращило стабільність навчання в цілому. Використання швидкості навчання більше за 1e-4 призводило до нестабільної мінімізації функції втрат.

Навчання генеративної моделі: Для перевірки гіпотез були навчені дев’ять генеративних моделей на кожну комбінацію розміру патчів (2x2, 4x4, 8x8 пікселів) та ступені маскування (20/40/60 відсотків). На Рис. 2 зображені приклади маскування вхідного зображення.

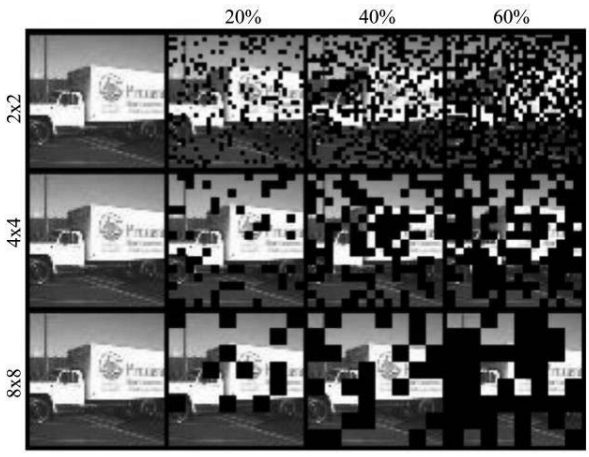

Рис. 2. Приклади маскування вхідного зображення (колонка 1) в залежності від розміру патча та відсотку маскування.

Навчання класифікатора: Кожна варіація генеративної моделі була навчена на задачу класифікації використовуючи доданий шар після токену класу. Додатково була навчена модель класифікації без попереднього навчання генеративної моделі (базова модель) та використовувалась для порівняння з іншими.

Пов'язані роботи

Методи передавального навчання завжди покращують результати навчання нейронних мереж. Модель UVCGANv2 використовує такий самий метод маскування, як і в цій роботі, для покращення результату перетворення зображень, проте ця модель не вирішує задач класифікації. Модель ViT попередньо навчалася на більших даних для покращення результатів класифікації на вибірці даних ImageNet. Метод, запропонований в BEIT [1], використовує ту саму вибірку даних для попереднього навчання, але маскує візуальні токени та використовує окрему мережу для побудови словника токенів. Ще одного покращення ViT було досягнуто за допомогою використання додаткових згорткових нейронних шарів [13].

У цій роботі випробувано метод, який використовує ту ж саму вибірку даних для попереднього навчання. Але маскування відбувається для пікселів зображення. Також модель UVCGANv2 має декілька згорткових нейронних шарів для перетворення вхідного зображення в візуальні токени.

Найбільш схожий метод з відновленням пікселів був застосований в SimMIM [14] – в ньому маскуються токени і за допомогою одного додаткового шару відбувається навчання відновлення пікселів замаскованого токену. В нашій роботі використовується генеративна мережа для відновлення всього зображення.

Варто зауважити, що проблема навчання візуального трансформера на малих даних була ефективно вирішена в ASTROFORMER [4] за допомогою модифікації архітектури трансформера та механізму самоуваги. Нова архітектура показала точність класифікації на даних Tiny ImageNet 92.98%, що значно перевищує результати цієї роботи. Має сенс спробувати метод, запропонований в цій роботі, після модифікації генеративної мережі використовуючи архітектуру ASTROFORMER.

Результати експериментів

Результати навчання генеративної моделі. Перевірка навчання генеративної моделі проводилася за рахунок порівняння значення функції втрат. Усі моделі не були спроможні деталізовано відтворити відсутні частини зображення, але задовільно зберігали форми присутніх на зображенні об'єктів. Як видно з результатів у Таблиці 1, гіпотеза щодо збільшення величини функції втрат у процесі збільшення розміру патчів підтвердилася.

Результати навчання класифікатора. Гіпотеза щодо покращення точності класифікації, використовуючи попередньо навчену генеративну модель, підтвердилася.

Вплив розміру патча та відсотка маскування на результат навчання класифікатора наведений в Таблиці 2. Неоднозначна кореляція точності класифікатора та конфігурації алгоритму маскування вхідного зображення потребують додаткових експериментів.

Розмір патчу 4x4 пікселів та ступінь маскування в 20 відсотків дало найкращий результат після подальшого навчання класифікатора. Точність класифікації збільшилась на 4.7 процентних пункти вище за базову модель (з 40% до 44%).

Таблиця 1

Функція втрат генеративної моделі після навчання для кожної комбінації розміру патча та відсотку маскування

Розмір патчу / Відсоток маскування	20%	40%	60%
2x2	0.0178	0.0309	0.0445
4x4	0.0185	0.0344	0.0521
8x8	0.0224	0.0403	0.0646

Таблиця 2

Точність класифікації для базової моделі та генеративних після навчання на задачу класифікації

Модель	Точність класифікації
базова модель	40%
розмір патчу 2x2	
20% маскування	**44.6**
40% маскування	44.1
60% маскування	**44.6**
розмір патчу 4x4	
20% маскування	**44.7**
40% маскування	43.2
60% маскування	43.6
розмір патчу 8x8	
20% маскування	43.1
40% маскування	**44.3**
60% маскування	44.1

Зміна точності передбачень під час навчання для валідаційних даних показана на Рис. 3, 4, 5 для кожного розміру патча та ступеня маскування.

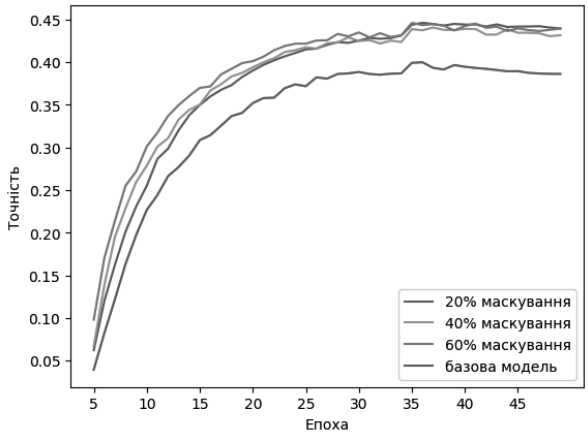

Рис. 3. Патчі розміром 2x2.

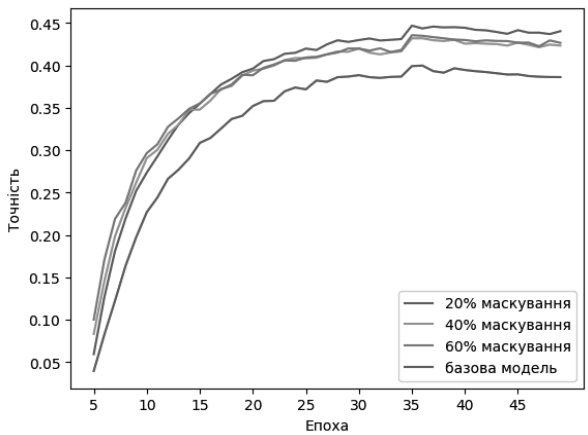

Рис. 4. Патчі розміром 4x4.

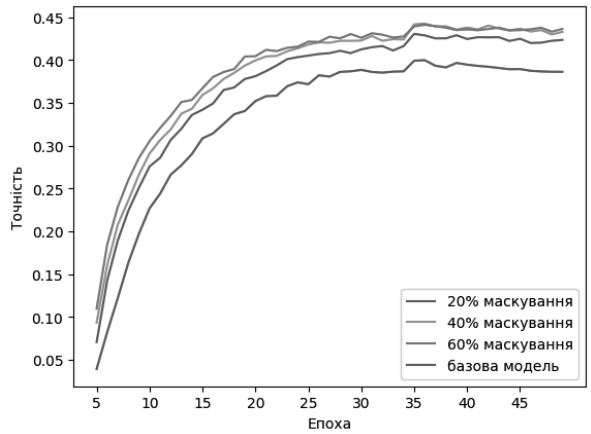

Рис. 5. Патчі розміром 8x8.

Висновки

Виконано серію експериментів з попереднього навчання модифікованої генеративної моделі UVCGANv2 і подальшого передання знань для навчання моделі класифікації зображень. Навіть на невеликих даних запропонований метод покращив точність класифікації на 4.7%.

Дослідження впливу розмірності патчів для маскування зображення, а також загального ступеня маскування, підтвердило гіпотезу про те, що збільшення розміру патчів збільшує значення функції втрат у процесі відновлення зображення генеративною моделлю.

Результати отримані в цій роботі доводять потенціал запропонованого методу. Предметом уваги подальшого дослідження цього методу можуть стати використання більшої вибірки даних для попереднього тренування, використання більшого розміру вхідного зображення, зміна архітектури UVCGANv2 на запропоновану в ASTROFORMER, та використання методів регуляризації.

References

1. Bao H., Dong L., Piao S. and Wei F. (2021) BEiT: BERT Pre-Training of Image Transformers, arXiv preprint arXiv:2106.08254.
2. Bachlechner T., Majumder B.P., Mao H.H., Cottrell G.W. and McAuley J. (2021) ReZero is All You Need: Fast Convergence at Large Depth, Uncertainty in Artificial Intelligence. PMLR.

3. Cubuk E.D., Zoph B., Mane D., Vasudevan V. and Le Q.V. (2019) AutoAugment: Learning Augmentation Strategies from Data Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
4. Dagli R., (2023) Astroformer: More Data Might Not be All You Need for Classification, arXiv preprint arXiv:2304.05350
5. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv preprint arXiv:2010.11929.
6. He K., Zhang X., Ren S. and Sun J. (2016) Deep Residual Learning for Image Recognition, Proceedings of the IEEE conference on computer vision and pattern recognition.
7. Kingma D.P., Ba J. (2014) Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980.
8. Ronneberger O., Fischer P. and Brox T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation, Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing.
9. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V. and Rabinovich A. (2015) Going Deeper with Convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition.
10. Tan M., Le Q.V. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, International conference on machine learning. PMLR.
11. Torbunov D., Huang Y., Tseng H., Yu H., Huang J., Yoo S., Lin M., Viren B. and Ren Y. (2023) UVCGAN v2: An Improved Cycle-Consistent GAN for Unpaired Image-to-Image Translation, arXiv preprint arXiv:2303.16280.
12. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L. and Polosukhin I. (2017) Attention Is All You Need, Advances in neural information processing systems 30.
13. Xiao T., Singh M., Mintun E., Darrell T., Dollár P. and Girshick R. (2021) Early Convolutions Help Transformers See Better, Advances in neural information processing systems 34.
14. Xie Z., Zhang Z., Cao Y., Lin Y., Bao J., Yao Z., Dai Q., Hu H. (2022) SimMIM: A Simple Framework for Masked Image Modeling, International Conference on Computer Vision and Pattern Recognition (CVPR)

Одержано: 09.04.2024
Внутрішня рецензія отримана: 21.04.2024
Зовнішня рецензія отримана: 26.04.2024

Про авторів:

[1]*Терентьєв Роман Валерійович*,
магістр

[1]*Іваненко Павло Андрійович*,
кандидат фізико–математичних наук,
старший науковий співробітник.
https://orcid.org/0000-0001-5437-9763.

Місце роботи авторів:

[1]Інститут програмних систем
НАН України,
тел. +38-044-522-62-42
E-mail: ukrprog@isofts.kiev.ua
www.iss.nas.gov.ua