Обзор-рецензия на статью Exploring Plain Vision Transformer Backbones for Object Detection (Y. Li et al.)

Колодезного Александра
ФКН БПМИ192

27 сентября 2022 г.

1 Содержание работы

Работа посвящена разработке архитектуры ViTDet (Y. Li, Mao et al. (2022)) для детектирования объектов на изображении (задача Object Detection). При этом в качестве нейросети для подсчёта карт признаков, называемой в статье backbone, стремятся использовать стандартную архитектуру трансформера для изображений ViT (Dosovitskiy et al. (2020)) предобученную без учителя с минимальными изменениями и насчитанные карты признаков использовать в стандартных приёмах детектирования, таких как Mask R-CNN (He, Gkioxari et al. (2017)).

Стандартная архитектура трансформера предполагает одно и то же разрешение изображения на всех слоях, за счёт чего невозможно применить стандартный приём FPN, однако как выясняется в статье использование простой пирамиды признаков (simple feature pyramid) достаточно, чтобы модель показывала результат сравнимый с современными моделями.

2 История работы

Первая версия работы была опубликована в мае 2022 года, группой Facebook AI research за авторством Yanghao Li, Hanzi Mao, Ross Girshick,

Kaiming He. Эта группа исследователей активно занималась проблемой детектирования изображений и до опубликованной статьи. Некоторые из них разрабатывали такие методы в сфере детектирования изображений как R-CNN (Girshick et al. (2014)), Fast R-CNN (Girshick (2015)), Faster R-CNN (Ren et al. (2015)), Mask R-CNN (He, Gkioxari et al. (2017)), так же ими были предложен метод обучения трансформера изображений Mask Autoencoders (MAE He, X. Chen et al. (2022)) и архитектура детектирования изображений с использованием трансформера (MViTv2 Y. Li, Wu et al. (2022)). Данная работа во многом является эволюционным продолжением представленной линейки архитектур, в которой комбинируются разные уже существующие методы.

Если говорить про недавние работы, которые непосредственно повлияли на данную, то стоит выделить работы связанные с применением трансформеров для задач Object detection. Наиболее примечательная это Swin (Ze Liu et al. (2021)) от Microsoft Research Asia, вышедшая в марте 2021 года, которая имеет иерархические слои трансформера, то есть слои уменьшаются в размерах. Параллельно с этой работой группа Facebook AI research опубликовала MViT (Fan et al. (2021)) в апреле 2021 г, которая так же имеет иерархическую структуру слоёв и далее улучшила её до MViTv2 (Y. Li, Wu et al. (2022)) для задачи Object Detection. Однако во всех этих работах сильно меняли архитектуру трансформера ViT и приходилось использовать для предобучения модели метод обучения с учителем. Однако в 2021 году были предложены методы обучения трансформеров изображений без учителя, такие как BEiT(Bao et al. (2021)) и MAE (He, X. Chen et al. (2022)), но они были предложены только для стандартного трансформера ViT и не применимы к иерархическим трансформерам Swin и MViT. Поэтому группа авторов нашей статьи делала попытки адаптировать обычный ViT трансформер с предобученным методом MAE в качестве backbone для архитектуры детектирования изображений, эти попытки представлены в неопубликованной предварительной статье, где предобученный ViT адаптируют в иерархическую структуру для использования с FPN(Lin et al. (2017)). Можно предположить что именно в процессе сравнения этого ViT с различными другими конфигурациями выяснилось, что предобученный трансформер хорошо работает и без FPN без адаптации под иерархическую структуру с использованием выхода только последнего слоя, то есть с минимальными изменениями. Однако для использования на картинках с высоким разрешением всё равно пришлось добавить некоторые модификации, но эти модификации позволяют использовать уже предобученные

веса.

3 Прямые конкуренты

Стоит отметить, что в декабре 2021 года была выложена похожая статья от группы из компании Google (W. Chen et al. (2021)). Идея статьи сильно совпадает с идеей рецензируемой статья, основанной на применении изначального трансформера ViT без дополнительных модификаций, в этой статье они назвали свою архитектуру UViT (Universal Vision Transformer).

Отличие двух работ связанно скорее с использованием различных приёмов и проведением различных экспериментов. В статье UViT изучают различные конфигурации стандартного ViT трансформера, такие как грубина сети, длина векторов представления, используемое разрешение изображений, при этом для детектирования используется только последний слой трансформера, в то время как в статье про VitDet сосредоточились на ViT с широко используемыми параметрами, но при этом добавляют дополнительные модификации, такие как простая пирамида признаков (Simple Feature Pyramid) и использование свёрточных слоёв, а так же активно используется метод обучения без учителя и в результате получили значительно лучшее качество.

Весьма вероятно, что так как работа про UVit была опубликована после выкладывания предварительной работы и до публикования основной, то она внесла важный вклад в развитие работы.

4 Последующие работы

На данный момент есть по разным сайтам 20-30 статей цитирующие данную работу.
Прямого продолжения данной работы от той же группы авторов нет, однако другие исследователи пытаются в дальнейшем улучшать архитектуру. Одним из направлений является попытка вместе с предобученной без учителя моделью ViT использовать остающийся после метода MAE декодер (сам ViT является энкодером в метода MAE), данный подход разрабатывается в двух работах Fang et al. (2022) и Zhang et al. (2022).

5 Характеристика работы

Учитывая, что в последнее время трансформеры активно адаптируют под различные задачи компьютерного зрения с различными модификациями архитектуры, вопрос построения некоторой базовой модели, с которой будут сравниваться более сложные модели действительно является актуальным, и данная статья во многом закрывает данный вопрос предлагая сильные метрики при применении базовой модели ViT. И хотя как можно видеть авторы работы не первые, кто пытается ответить на данный вопрос, но именно у них получилось качества сравнимого с наиболее продвинутыми на сегодняшний день моделями.

Если не разбираться базово в методах решения задач, компьютерного зрения, то по тексту работы не будет понятно с первого раза, однако если знать классические методы решения задачи детекции изображений, то читатель быстро поймёт смысл работы, так как он написан довольно ясным языком и логично построен.

Что касается экспериментов, то для каждой части архитектуры, например для использования простой пирамиды признаков, для использования window attention и других, поставлен эксперимент проверяющий необходимость данного метода и сравнивающий его с другими вариантами реализации.

Стоит также отметить, что в работе полностью описаны детали реализации и используемые гиперпараметры для обучения каждой их сравниваемых моделей, что теоретически позволяет воспроизвести результат статьи. Однако слабой стороной статьи является недостаточно сравнение с SOTA моделями, а именно в работе сравниваются только со статьями Swin и MViTv2 с несколькими фреймворками Mask R-CNN и Cascade Mask R-CNN, перебрав гиперпараметры для Swin и MViTv2. Однако в результате значения для этих моделей в статье ViTDet хуже чем в оригинальных статьях про эти модели, что может ввести в заблуждение. Во многом это связано с тем, что например для Swin в оригинальной статье использовался другую голову для детекции HTC++ (K. Chen et al. (2019)).

Также в статье стоило бы провести дополнительные эксперименты с другими методами детекции, например с HTC++, на котором у Swin лучшее качество. Также можно было бы сравнить другие методы обучения без учителя, такие как BEiT (Bao et al. (2021)).

Список литературы

1. Hangbo Bao, Li Dong и Furu Wei. "Beit: Bert pre-training of image transformers". в: *arXiv preprint arXiv:2106.08254* (2021).
2. Kai Chen et al. "Hybrid task cascade for instance segmentation". в: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, с. 4974—4983.
3. Wuyang Chen et al. "A simple single-scale vision transformer for object localization and instance segmentation". в: *arXiv preprint arXiv:2112.09747* (2021).
4. Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". в: *arXiv preprint arXiv:2010.11929* (2020).
5. Haoqi Fan et al. "Multiscale vision transformers". в: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, с. 6824—6835.
6. Yuxin Fang et al. "Unleashing vanilla vision transformer with masked image modeling for object detection". в: *arXiv preprint arXiv:2204.02964* (2022).
7. Ross Girshick. "Fast r-cnn". в: *Proceedings of the IEEE international conference on computer vision*. 2015, с. 1440—1448.
8. Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". в: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, с. 580—587.
9. Kaiming He, Xinlei Chen et al. "Masked autoencoders are scalable vision learners". в: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, с. 16000—16009.
10. Kaiming He, Georgia Gkioxari et al. "Mask r-cnn". в: *Proceedings of the IEEE international conference on computer vision*. 2017, с. 2961—2969.
11. Yanghao Li, Hanzi Mao et al. "Exploring plain vision transformer backbones for object detection". в: *arXiv preprint arXiv:2203.16527* (2022).
12. Yanghao Li, Chao-Yuan Wu et al. "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection". в: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, с. 4804—4814.

13. Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. в: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, с. 2117—2125.

14. Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. в: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, с. 10012—10022.

15. Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. в: *Advances in neural information processing systems* 28 (2015).

16. Xiaosong Zhang et al. “Integral Migrating Pre-trained Transformer Encoder-decoders for Visual Object Detection”. в: *arXiv preprint arXiv:2205.09613* (2022).