Городилова Анастасия

Рецензия:

A ConvNet for the 2020s

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie

1. Работа авторов представляет из себя последовательное улучшение модели ResNet50 (одной из самых мощных сверточных сетей); авторы применяли поочередно различные изменения в архитектуре, которые используются в Swin Transformer (наиболее мощная версия визуального трансформера). Таким образом, авторы обогнали результаты Swin Transformer на ImageNet-1K, COCO, ADE20K. Итоговое множество сверточных моделей (которые отличаются лишь количеством обучаемых параметров) авторы называют ConvNext, и их особенностью является использование характерных для архитектуры трансформер операций, за исключеним механизма внимания.

2. Работы выпущена (на ArXiv) в январе 2022 года, а этим летом была представлена на топовой конференции по компьютерному зрению CVPR. Авторы работы: Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie. Авторы выпустили работу в процессе стажировки в Facebook AI Research. Многие из авторов давно работают в области компьютерного зрения (и в области исследований сверточных сетей и трансформеров), и имеют в ней множество статей; более того, многие идеи из статьи они уже исследовали. К примеру, Zhuang Liu был автором DenseNet, одной из популярных сверточных сетей; Christoph Feichtenhofer исследовал масштабируемый визуальный трансформер в контексте комбинированного использования сверточной архитектуры и архитектуры трансформер (Multiscale vision transformers. ICCV, 2021). Trevor Darrell в 2014 году исследовал задачу сегментации, которую затрагивают авторы рассматриваемой статьи (Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014), и также исследовал свертки в трансформерах (Early convolutions help transformers see better. In NeurIPS, 2021). Saining Xie был автором нейросети ResNext (Aggregated residual transformations for deep neural networks. In CVPR, 2017), и именно ее авторы используют в качестве промежуточного улучшения.

3. Базовыми статьями для работы являются ResNet & SwinTransformer: авторы берут первую архитектуру, и поэтапными изменениями приводят первую архитектуру ко второй, не используя Attention.

Помимо уже упомянутого ResNeXt, важными промежуточными решениями являются depth-wise свертки, которые пришли из моделей MobileNet, Xception. Также, важными статьями являются "Revisiting resnets: Improved training and scaling strategies. NeurIPS, 2021" и "Resnet strikes back: An improved training procedure in timm", в которых авторы применили к сети ResNet улучшенную процедуру обучения и достигли повышения результатов - именно с этого начинается и рассматриваемая статья, где авторы копируют процедуру обучения у DeiT и SwinTransformer.

4. У статьи очень много цитирований: 275, и большая часть из них объясняется тем, что архитектура ConvNext всего за несколько месяцев стала очень популярной, и используется в самых разных задачах. Например, в статье "VidConv: A modernized 2D ConvNet for Efficient Video Recognition, 2022" как часть архитектуры модели для работы с видео авторы использовали ConvNeXt . Одним из основных продолжений / конкурентов является статья "More ConvNets in the 2020s: Scaling up Kernels Beyond 51 × 51 using Sparsity, 2022". Статья вышла спустя полгода после обозреваемой, и развила парадигму улучшения сверточных сетей; авторы использовали ядра сверток размера 51 * 51, и обогнали как SWIN, так и ConvNeXt.

5. -

6. Очевидным плюсом статьи является конкурентность разработанной модели: она обгоняет на различных задачах SwinTransformer при схожем количестве требуемых вычислений, и уже используется в различных применениях. Также, в принципе подобную тематику интересно изучать: ведь на CVPR 2022 года большинство статей посвящено работе с трансформерами и решению гораздо более сложных задач, чем классические CV задачи по типу классификации и сегментации. Здесь же авторы делают общее исследование, которое пытается вывести промежуточное состояние между двумя эпохами компьютерного зрения; любые такие исследования, которые в целом обозревают предметную область, выделяются из остальных.

7. Я вижу единственным минусом статьи то, что авторы никак не сворачивают с первоначального плана: взять ResNet и последовательно добавлять те приемы, которые использует Swin Transformer. Дело в том, что около половины этих приемов никак не улучшают модель (либо качество остается тем же, либо растет на 0.1%). Авторы не исследовали, почему так происходит; что будет, если

эти приемы не делать (ведь зачем они нужны, если не улучшают качества)

8. Было бы интересно рассмотреть другие сверточные архитектуры, такие как VGG, MobileNet, и др., и применить к ним подобную процедуру перехода к Swin Transformer.

9. Нейросети очень неустойчивы к adversarial-attacks; нейросеть очень легко обмануть, например, немного поменять исходную картинку так, что класс, выдаваемый моделью, очень сильно изменится. Интересно, что чем тяжелее модель, тем проще ее обмануть: и результаты обмана трансформер-моделей, как моделей очень тяжелых, еще более абсурдны, чем у сверточных сетей. Интересно исследовать на устойчивость к таким атакам авторскую модель: ведь она является чем-то промежуточным между сверточными сетями и трансформерами.