Отчет рецензента-исследователя

Макоян Артем Каренович

1. В работе вводится новый способ внедрения механизма внимания в задачи компьютерного зрения. Удается получить **линейное** от количества пикселей время работы, что позволяет рассматривать блоки внимания как новый универсальный скелет для задач компьютерного зрения, превосходящий CNN.
2.
    a. Дата: 17 августа 2021
    b. Конференция ICCV 2021 (best paper award)
    c. Авторы: Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo
    d. Компания: Microsoft Research Asia
    e. Схожие работы авторов: ???
    f. Данная работа - это непосредственное улучшение “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE”.
3. Базовые статьи:
    a. AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE
    b. См заметки
4. Swin Transformer V2: Scaling Up Capacity and Resolution - прямое продолжение
5. Прямые конкуренты: см заметки
6. Сильные стороны
    a. Актуальность (трансформеры довольно востребованная архитектура для зрения).
    b. Хороший код, предоставлены репозитории под каждую задачу КЗ
    c. Востребованный на практике подход
    d. Хороший Ablation Study
7. Слабые стороны: слабых сторон не нашел.
8. Очевидно, что текущий скелет будет улучшать работу сеток во многих задачах компьютерного зрения, а учитывая его небольшую асимптотику работы, его можно будет применять во многих задачах. Хотелось бы более подробно сравнить данный подход для внедрения трансформеров с подходом предыдущих работ.

Заметки:

- Sliding window based self-attention approaches [33, 50]
- Self-attention based backbone architectures [33, 50, 80]
- Self-attention/Transformers to complement CNNs [67, 7, 3, 71, 23, 74, 55], head networks [32, 27]
    - Transformer based vision backbones: Vision Transformer (ViT) [20] and its follow-ups [63, 72, 15, 28, 66]
- Concurrent to our work are some that modify the ViT architecture [72, 15, 28]
- Another concurrent work [66] explores a similar line of thinking to build multi-resolution feature maps on Transformers. Its complexity is still quadratic to image size, while ours is linear and also operates locally