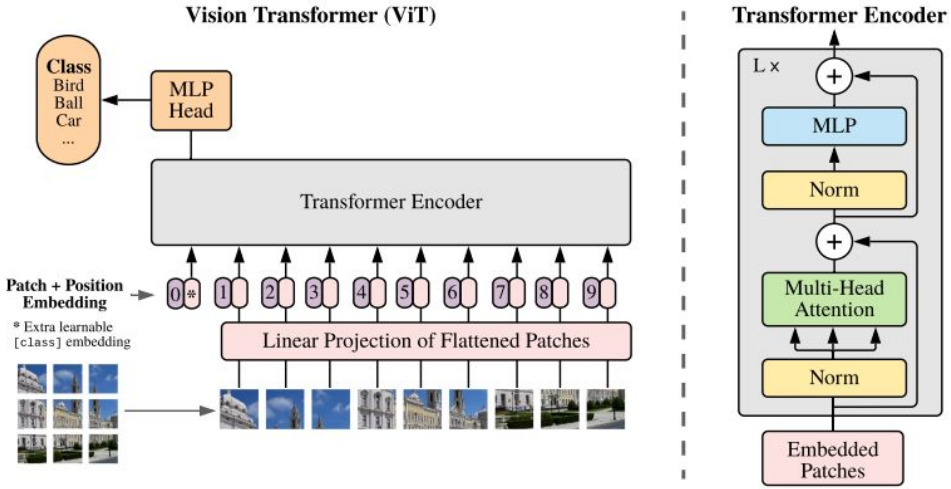
Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Подготовил:
Макоян Артем Каренович, БПМИ 192

SWIN

- 17 августа 2021 года, ICCV 2021, best paper award
- Авторы: Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo - Microsoft Research Asia

Работа - это **логическое продолжение** “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE”



SWIN. Основной вклад.

Главная цель работы: *разработать универсальный backbone на основе трансформеров*

- Архитектуру можно будет легко внедрить в любую задачу компьютерного зрения
- Скорость работы должна быть приемлемой - линейная от размера картинки
- Кривая трейд-оффа (между скоростью и качеством) лучше, чем у CNN и других конкурентов

SWIN. Предшественники.

Раньше всех были *CNN*.

- AlexNet [39]
- VGG [52]
- GoogleNet [57]
- ResNet [30]
- DenseNet [34]
- HRNet [65]
- EfficientNet [58]

Проблемы: трансформеры получают качество лучше

SWIN. Предшественники.

Self-attention ***вместо конволюций***. В базовых CNN архитектурах заменяем слой конволюций на self-attention блоки.

- Local relation networks for image recognition [33]
- Stand-alone self-attention in vision models [50]
- Exploring self-attention for image recognition [80]

Проблемы: доступ к памяти очень дорогой, из-за этого при одинаковом кол-ве FLOPs конволюции намного быстрее (так как хорошо оптимизированы)

SWIN. Предшественники.

*Self-attention/Transformers **дополняющие** CNN*. Аугментируем конволюции механизмом внимания, по сути encoder-decoder архитектура, где encoder - трансформер. SWIN - продолжение исследований данной идеи.

- Non-local neural networks [67]
- Gcnet: Non-local networks meet squeeze-excitation networks and beyond[7]
- Attention augmented convolutional networks [3]
- Disentangled non-local neural networks [71]
- Dual attention network for scene segmentation [23]
- Ocnet: Object context network for scene parsing [74]
- Bottleneck transformers for visual recognition [55]

Проблемы: результаты хуже SWIN, сетки достаточно тяжелые, квадратичная сложность

SWIN. Предшественники.

Скелеты, основанные на трансформерах. Продолжение мысли ViT, как и SWIN. По сути прямые **конкуренты**.

- Vision Transformer (ViT) [20]
- DeiT [63] (ViT, но для обучения нужно меньше данных)
- Tokens-to-token vit: Training vision transformers from scratch on imagenet [72]
- Do we really need explicit position encodings for vision transformers? [15]
- Transformer in transformer [28]
- Pyramid vision transformer: A versatile backbone for dense prediction without convolutions [66]
- Toward transformer-based object detection [2]

Проблемы: результаты хуже SWIN, квадратичное время работы

Swin Transformer V2: Scaling Up Capacity and Resolution

Прямое продолжение от тех же авторов. Добавили несколько трюков для более стабильного обучения и уменьшения потребности в большом количестве данных. Также в статье было подмечено, что для оригинального SWIN:

- An instability issue when scaling up model capacity
- Degraded performance when transferring models across window resolutions

Как раз эти проблемы решает новая архитектура.

SWIN. Сильные стороны работы

- *Значительные результаты для науки и индустрии.*
- *Актуальная тема*. Заслуженный ажиотаж вокруг трансформеров.
- *Хорошая и удобная кодовая база*. Много примеров для разных задач компьютерного зрения.
- *Ablation study*. Понятно насколько критичны трюки из статьи: Shifted windows, Relative position bias, сравнение методов self-attention
- Приведено подробное сравнение с конкурентами и SotA: ResNe(X)t, DeiT и т.д.

SWIN. Слабые стороны работы

В принципе каких-то больших минусов не обнаружили, но не хватает понимания насколько хорошо работает модель на **маленьком количестве данных**.

Видимо, как и ViT, SWIN требует большое количество данных.

В этом же направлении можно провести дальнейшие исследования, попробовать перенести трюки из DeiT для более стабильной работы.