Swin Transformer and 5 Reasons to Use Transformer/Attention in Computer Vision

Han Hu
Microsoft Research Asia
June 20th, 2021

CVPR21, The 3rd Tutorial on “Learning Representations via Graph-structured Networks”
What is the role of Transformer for computer vision?

Transformer (2017.6)
An answer: will also refresh & dominate CV

ImageNet-1K image classification

Mostly ViT

CNN backbones

Transformer backbones
Vision Transformer (ViT, 10/2020)

- SOTA performance on Image classification

Alexey Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR’ 2021
An answer: will also refresh & dominate CV

COCO object detection

ADE20K semantic segmentation

CNN backbones
Transformer backbones

Swin Transformer

Swin Transformer

Other models
Models with highest box AP

Other models
Models with highest mIoU
Swin Transformer (03/2021)

- SOTA performance on object detection and semantic segmentation

Transformer
(strong modeling power)

good priors for visual signals
(hierarchy / locality / translation invariance)

4 years unleash the power of Transformer in CV

Reason I: General modeling capability
Reason II: Complement convolution
Reason III: Strong modeling power
Reason IV: Better connect vision and language
Reason V: Scalability

Swin Transformer: a general-purpose backbone
Reason I to use Transformer in computer vision

• General modeling capability
  • All concepts (concrete or abstract) and their relationships can be modeled by a graph
  • Modeling arbitrary relationship via verification, which is hard by CNN
Reason I to use Transformer in computer vision

• General modeling capability
  • Can model all of pixel-to-pixel, object-to-pixel, object-to-object relationships
Relation Networks for Object Detection (CVPR’2018)

It is much easier to detect the *glove* if we know there is a *baseball player*. 
Relation Networks for Object Detection (CVPR’2018)

Key Feature
✓ Relative position

Han Hu et al. *Relation Networks for Object Detection*. CVPR 2018
Relation Networks for Object Detection (CVPR’2018)

• The first fully end-to-end object detector

Han Hu et al. Relation Networks for Object Detection. CVPR 2018
DeTR (ECCV’2020)

• Another end-to-end object detector
Reason II to use Transformer in computer vision

- Complement convolution
  - “Convolution is too local!”
  - Global (Transformer) vs. local (conv.)
Non-local networks (CVPR’2018)
The Degeneration Problem of NLNet

• Expectation of Ideally Learnt Relation
  • Different queries affected by different key
The Degeneration Problem of NLNet

• What does the Self-Attention Learn?
  • Different queries affected by the same keys
Visualizations on Real Tasks

- ⊘ indicates the query point
- The activation map for different queries are similar
- The self-attention model degenerates to a unary model

[GCNet, ICCVW'2019]
GCNet (ICCVW’2019, PAMI’2021)

- Find the degeneration issue in computer vision
- Explicitly leverage degenerated formulation for better efficiency

<table>
<thead>
<tr>
<th></th>
<th>Non-Local Block</th>
<th>Global Context Block</th>
<th>Reduction ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLOPs</td>
<td>9.3G</td>
<td>4.0M</td>
<td>2,300x</td>
</tr>
<tr>
<td>model size</td>
<td>2.1M</td>
<td>0.1M</td>
<td>20x</td>
</tr>
<tr>
<td>accuracy (mAP)</td>
<td>38.0</td>
<td>38.1</td>
<td>unchanged</td>
</tr>
</tbody>
</table>
Disentangled non-local networks (ECCV’2020)

- Solve the degeneration problem
Reason III to use Transformer in computer vision

- Powerful due to **adaptive computation**
- “Convolution is exponentially inefficient!”
Local relation networks (2019.4)

- Transformer as backbones

<table>
<thead>
<tr>
<th>Stage</th>
<th>Output</th>
<th>ResNet-50</th>
<th>LR-Net-50 (7×7, m=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>res1</td>
<td>112×112</td>
<td>7×7 conv, 64, stride 2</td>
<td>1×1, 64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3×3 max pool, stride 2</td>
<td>7×7 LR, 100</td>
</tr>
<tr>
<td>res2</td>
<td>56×56</td>
<td>1×1, 64</td>
<td>1×1, 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3×3 conv, 64</td>
<td>7×7 LR, 200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1×1, 256</td>
<td>1×1, 256</td>
</tr>
<tr>
<td>res3</td>
<td>28×28</td>
<td>1×1, 128</td>
<td>1×1, 200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3×3 conv, 128</td>
<td>7×7 LR, 400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1×1, 512</td>
<td>1×1, 512</td>
</tr>
<tr>
<td>res4</td>
<td>14×14</td>
<td>1×1, 256</td>
<td>1×1, 400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3×3 conv, 256</td>
<td>7×7 LR, 800</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1×1, 1024</td>
<td>1×1, 1024</td>
</tr>
<tr>
<td>res5</td>
<td>7×7</td>
<td>1×1, 512</td>
<td>1×1, 800</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3×3 conv, 512</td>
<td>7×7 LR, 800</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1×1, 2048</td>
<td>1×1, 2048</td>
</tr>
<tr>
<td>1×1</td>
<td>global average pool</td>
<td>1×1, 2048</td>
<td>1×1, 2048</td>
</tr>
</tbody>
</table>

- # params: \(25.5 \times 10^6\) for ResNet, \(23.3 \times 10^6\) for LR-Net
- FLOPs: \(4.3 \times 10^9\) for ResNet, \(4.3 \times 10^9\) for LR-Net

*Han Hu et al. Local Relation Networks for Visual Recognition. ICCV 2019*
But ... slow in real computation

- Because different queries use different key sets
Vision Transformer (ViT)

• by Google Brain (2020.10)

---

**Computation speed**

<table>
<thead>
<tr>
<th>Model</th>
<th>FLOPs</th>
<th>Speed (TPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R50</td>
<td>4.3G</td>
<td>~2100 im/s</td>
</tr>
<tr>
<td>ViT-B/32</td>
<td>4.3G</td>
<td>~3000 im/s</td>
</tr>
</tbody>
</table>

+50% speed-up

---

Alexey Dosovitskiy et al. an Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR’ 2021
Swin Transformer =

- Transformer
  - Strong modeling power
- + good priors for visual modeling
  - Hierarchy
  - Locality
  - Translational invariance

Transformer (ViT) → Swin Transformer

Patch/Feature bin
Computation scope of self-attention
Hierarchy

- Processing objects of different scales
Locality by non-overlapped windows

• Proves beneficial in modeling the high correlation in visual signals (Yann LeCun)
• Linear complexity with increasing image resolution: from $O(n^2)$ to $O(n)$

ViT: $256^2=65536$ (Global)
Swin Transformer: $16 \times 16^2=4096$ (Local)
Locality by non-overlapped windows

• Compared to sliding window (LR-Net)
  • Shared key set enables friendly memory access and is thus good for speed (larger than 3x)

sliding window (LR-Net)

Non-overlapped window (Swin Transformer)
Shifted non-overlapped windows

• Enable cross-window connection
  • Non-overlapped windows will result in no connection between windows
  • Performs as effective or even slightly better than the sliding window approach, due to regularization effects
Translational semi-invariance

- Relative position bias plays a more important role in vision than in NLP

\[
\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d} + B)V,
\]

semi-invariance is as effective as full-invariance in our experiments
Architecture instantiations

- Resolution of each stage is set similar as ResNet, to facilitate application to down-stream tasks
Application: object detection

- COCO object detection: #1 #2 #3 for single model (60.6 mAP)
  - Significantly surpass all previous CNN models (+3.5 mAP)

- COCO instance segmentation: #1 for single model (52.4 mAP)
  - Significantly surpass all previous CNN models (+3.3 mAP)
Application: object detection

- Performs consistently better than CNN on various object detectors and various model sizes (+3~4.5 mAP)
Application: semantic segmentation

- ADE20K semantic segmentation: \textbf{#1} for single model (53.9 mIoU)
  - The largest and most difficult semantic segmentation benchmark
    - 20,000 training images, 150 categories
  - Significantly surpass all previous CNN models (+5.5 mIoU vs. the previous best CNN model)
Application: video recognition (coming soon)

Figure 2: Overall architecture of Video Swin Transformer (tiny version, referred to as Swin-T).
Application: video recognition

- Swin Transformer achieves SOTA on major video benchmarks with 20x less pre-training data and 3x smaller model size

+2.9% using the same pre-training data

+3.6% using the same pre-training data
Reason IV to use Transformer in computer vision

• Better connect vision and language: unified modeling

CLIP
https://openai.com/blog/clip/
Reason V to use Transformer in computer vision

- Scalable to large model and large data

ViT G/14
- 1000 G Flops
- 2B parameters
- 3B images

Summary: 4 years unleash the power of Transformer in CV

Reason I: General modeling capability
- RelationNet (CVPR’2018)
- LearnRegionFeat (ECCV’2018)
- STRN (ICCV’2019)
- MEGA (CVPR’2020)
- DeTr (ECCV’2020)
- RelationNet++ (NeurIPS’2020)

2017.06
2017.11

Reason II: Complement convolution
- NLNet (CVPR’2018)
- GCNet (ICCVW’2019)
- DNL (ECCV’2020)

2019.4
2021.1

Reason III: Strong modeling power
- LRNet (ICCV’2019)
- ViT (ICLR’2021)

2021.6

Reason IV: Better connect vision and language
- CLIP (ICML’2021)

Reason V: Scalability
- ViT-G (Arxiv’2021)

Transformer (2017.6)