Transformers for medical image segmentation

SCAI-IDS Workshop 2023: AI and Medicine

THOME Nicolas – Prof. at SORBONNE University
ISIR Lab, MLIA TEAM

Loïc Themyr, Olivier Petit, Clément Rambour – Cnam Paris
Toby Collins, Alex Hostettler, IRCAD Strasbourg & Africa
Luc Soler, Visible Patient Inc
1. Context
2. U-Net Transformer
3. GLAM
Transformers everywhere since 2017

**NLP: BERT, GPT3, Chat-CPT, etc**

- Explain quantum theory to a child in the style of snoop dogg
  
  Yo, little homie, let me break it down for you
  Quantum theory, it's a crazy thing, it's true
  See, the world is made up of tiny, tiny particles
  That are always moving, and changing, and actin' wild

  These particles, they don't always behave like we think
  Sometimes they're waves, and sometimes they're points, it's a trip
  They can be in two places at once, and they can be in two states
  It's a mind-blowing concept, but that's the quantum fate

  So don't be scared, little homie, it's all good
  Quantum theory, it's just the way the world works, it's understood
  It's a wild ride, but it's a cool one too
  Quantum theory, it's the way the world does what it do.

**Vision since '21: Vision Image Transformer (ViT)**

Vision Transformer (ViT)
Transformers: self-attention

Key element: self-attention

- Each «token» re-embedded wrt all token
  \[ \Rightarrow \text{Global interactions} \]

- Self attention: \( O(N^2) \) complexity
  - Expensive (or impossible) for large \( N \)
Transformers for medical image segmentation

- Which features for medical image segmentation?
  - How to include long-range dependencies in U-shaped architectures
  - U-Transformer [MLMI’21]: self and cross attention in medical image segmentation
Transformers for medical image segmentation

- Full transformer architecture: more and more competitive in segmentation
- **How to adapt them to the** medical domain’s specificities?
  - 3D inputs: bottleneck in computational complexity
  - Ex: $512^3$ input volume: impossible to get tractable attention!
- **Full attention in 3D transformers with indirections (GLAM/FINE)** [WACV’23, MLMI’22]
1. Context
2. U-Net Transformer
3. Full attention in 3D transformers
Context

Organs segmentation illustration

Pancreas automatic segmentation
Context

Problems:
- High image resolution (512x512 pixels)
- Importance of global context in medical images segmentation
- Limitations of ConvNets receptive field

Abdominal CT-scan example showing receptive field's limitations
U-Net Transformer

Hybridation between U-Net [A] and Transformers

Segmentation example with U-Net's receptive field (red square) and U-Transformer’s attention map.

Architecture: U-Net Transformer

- Multi-Head Self-Attention (MHSA) in bottleneck (control complexity)
- Multi-Head Cross-Attention (MHCA) as an upsampling module
Architecture: Multi-Head Self-Attention

\[ X \in \mathbb{R}^{wh \times d}, W_q \in \mathbb{R}^{d \times d}, W_k \in \mathbb{R}^{d \times d}, W_v \in \mathbb{R}^{d \times d} \]
\[ Q = XW_q, K = XW_k, V = XW_v \]
\[ A = \text{Softmax}( \frac{QK^T}{\sqrt{d}} ) \]
\[ Y = AV \]
Architecture: Multi-Head Cross-Attention

**MHCA**: Filter high resolution features based on semantically richer features from the encoder.

\[ Y \text{ : Semantically richer features from bottleneck} \]

\[ S \text{ : High resolution features from skip connections} \]
Results

Experiments:

TCIA Pancreas dataset:
- Pancreas segmentation,
- 82 CT-scans
- 181~466 slices
- Size 512x512 pixels

Internal Multi-Organ dataset (IMO):
- 7 classes: liver, gallbladder, pancreas, spleen, right and left kidneys, stomach
- 85 CT-scans
- 57~500 slices
- Size 512x512 pixels
## Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TCIA</td>
<td>76.13 (± 0.94)</td>
<td>76.82 (± 1.26)</td>
<td>77.71 (± 1.31)</td>
<td>77.84 (± 2.59)</td>
<td><strong>78.50</strong> (± 1.92)</td>
</tr>
<tr>
<td>IMO</td>
<td>86.78 (± 1.72)</td>
<td>86.45 (± 1.69)</td>
<td>87.29 (± 1.34)</td>
<td>87.38 (± 1.53)</td>
<td><strong>88.08</strong> (± 1.37)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pancreas</td>
<td>69.71 (± 3.74)</td>
<td>68.65 (± 2.95)</td>
<td>71.64 (± 3.01)</td>
<td>71.87 (± 2.97)</td>
<td><strong>73.10</strong> (± 2.91)</td>
</tr>
<tr>
<td>Gallbladder</td>
<td>76.98 (± 6.60)</td>
<td>76.14 (± 6.98)</td>
<td>76.48 (± 6.12)</td>
<td>77.36 (± 6.22)</td>
<td><strong>78.32</strong> (± 6.12)</td>
</tr>
<tr>
<td>Stomach</td>
<td>83.51 (± 4.49)</td>
<td>82.73 (± 4.62)</td>
<td>84.83 (± 3.79)</td>
<td>84.42 (± 4.35)</td>
<td><strong>85.73</strong> (± 3.99)</td>
</tr>
<tr>
<td>Kidney(R)</td>
<td>92.36 (± 0.45)</td>
<td>92.88 (± 1.79)</td>
<td>92.91 (± 1.84)</td>
<td>92.98 (± 1.70)</td>
<td><strong>93.32</strong> (± 1.74)</td>
</tr>
<tr>
<td>Kidney(L)</td>
<td>93.06 (± 1.68)</td>
<td>92.89 (± 0.64)</td>
<td>92.95 (± 1.30)</td>
<td>92.82 (± 1.06)</td>
<td><strong>93.31</strong> (± 1.08)</td>
</tr>
<tr>
<td>Spleen</td>
<td>95.43 (± 1.76)</td>
<td>95.46 (± 1.95)</td>
<td>95.43 (± 2.16)</td>
<td>95.41 (± 2.21)</td>
<td><strong>95.74</strong> (± 2.07)</td>
</tr>
<tr>
<td>Liver</td>
<td>96.40 (± 0.72)</td>
<td>96.41 (± 0.52)</td>
<td>96.82 (± 0.34)</td>
<td>96.79 (± 0.29)</td>
<td><strong>97.03</strong> (± 0.31)</td>
</tr>
</tbody>
</table>
Results
Results

Ground Truth  Cross-attn level 1  Cross-attn level 2  Cross-attn level 3
1. Context
2. U-Net Transformer
3. Full attention in 3D transformers
Context

Inherent problem to high-resolution volumes segmentation:

- Size of the input
- Large memory requirements
- 180Gb for U-Net with image size 512x512x256

Common strategies to reduce the memory footprint:

- Limited model size
- Train on 2D slices
- Train on patches
  \[\Rightarrow \text{No long-range information}\]

- Downsampling
  \[\Rightarrow \text{Drop in quality}\]
Transformers for medical image segmentation

- **Transformers** became SOTA for **image segmentation** [B]
  - BUT: not possible to have full attention at high-resolution feature maps
- **Windowed transformers** [C,D] designed to reduce the complexity
  - BUT: no more long-range attention for high resolution feature maps

Windowed input at different hierarchy levels

Motivation

To keep the **full resolution** we work on patches:

Original image size: 512x512x256

Cropped patch size: **128x128x64**

Goal: learning a global representation of the full volume from batch training with crops
FINE: Full resolution mEmory transformer module

- Global representation embedded in **multiple level of memory tokens**
- **Visual tokens:** high-resolution 3D features (4x4x4)

- **Window token:** information at the window scale
- **Volume token:** learned representation of the full-size volume
Volume tokens

High-Resolution
Volume
Cropped patch

High-Resolution Volume

512x512x256 voxels

128x128x64 voxels

3D cropped patch

volume token  current volume token
Local window attention

512x512x256 voxels

128x128x64 voxels
Cross-window attention

[Diagram showing the process of cross-window attention with visual tokens, window tokens, volume tokens, and current volume tokens.]
Volume attention
Results

Synapse BCV [17]: CT scans Abdominal multi-organs segmentation
7 classes
30 volumes

Metrics:
- Dice score in % (DSC)
- 95% Hausdorff distance in mm (HD95)
Results
Results

Aorta

Ribs

Spine
Conclusion

**U-Net Transformer : MHSA + MHCA**

- Combination of powerful image segmentation models (U-Net) with long-range interaction model (Transformers)
- MHSA + MCASLong range interaction and spatial dependencies

**FINE :**

- Generic module for any windowed transformers
- Able to model **long-range** interaction beyond cropped input patch
- Perspective: hybrid FINE module in Conv architectures (nnU-Net)
Thank you