Medical image segmentation: transformer-based architectures and information flow

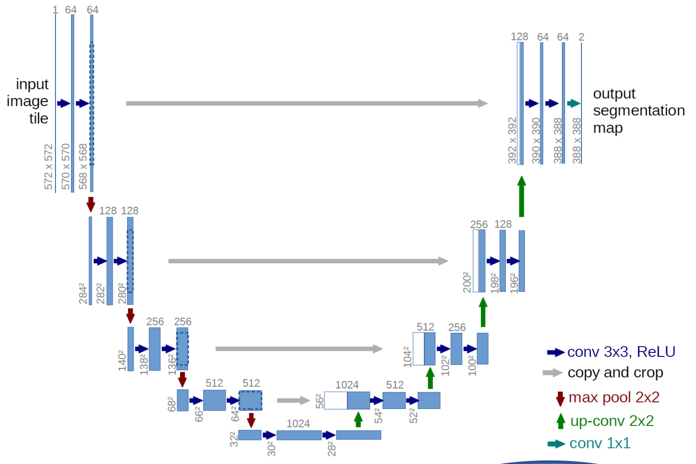
Caroline Petitjean

A joint work with S.N. Hasany and F. Mériaudeau

30 mars 2023

Segmentation in medical imaging

- The state-of-the-art model since 2015: UNet
  - Fully convolutional architecture
  - Variants: VNet, nnUNet, UNet++, etc



based on UNet

Generative models

GAN

GAN for image segmentation

DDPM

Diffusion for image segmentation

'14 '15 '16 '17 '18 '19 '20 '21

UNet

NLP transformer

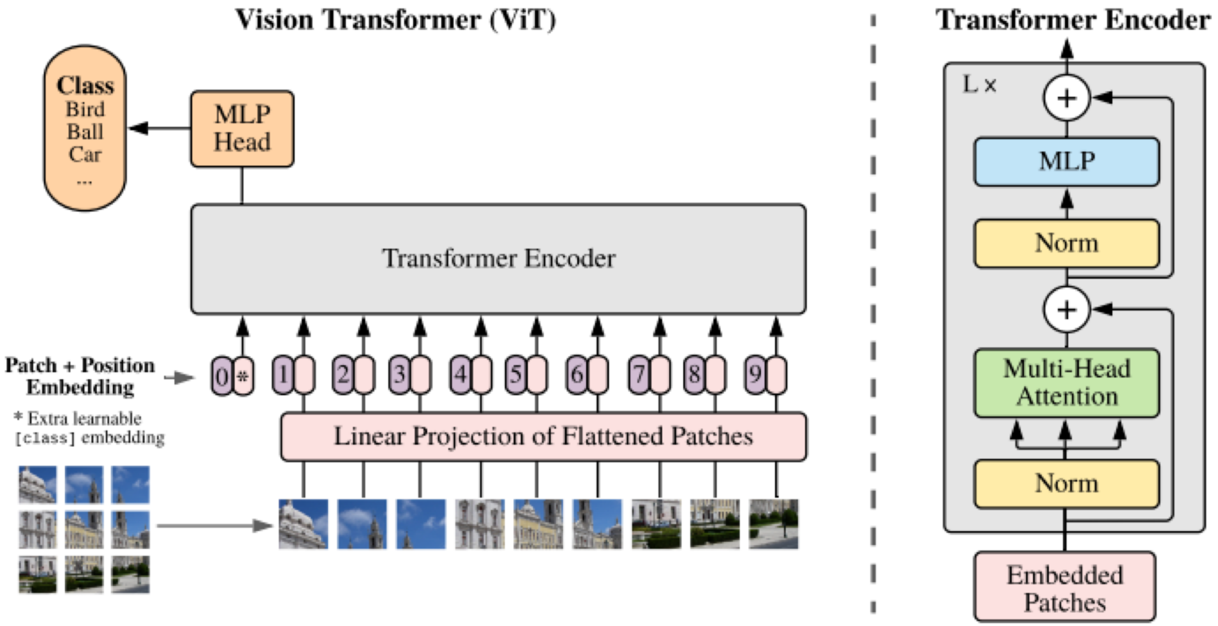
Vision transformer

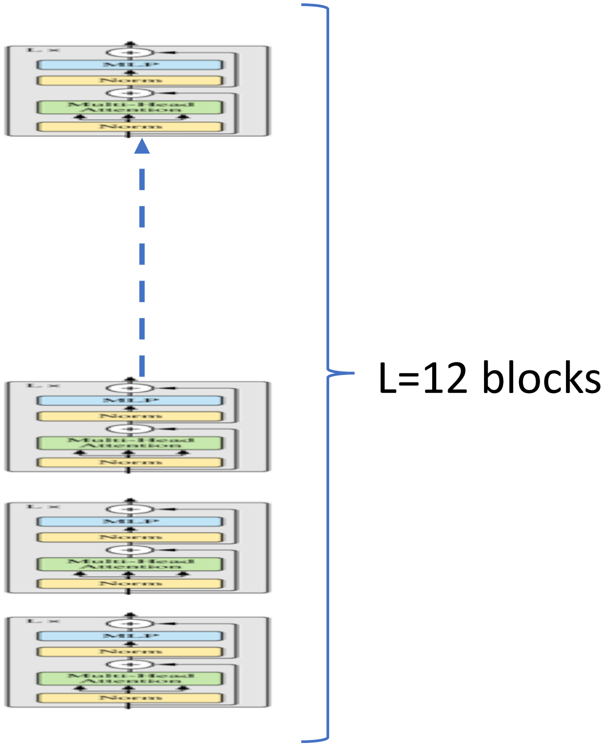
Transformer For image segmentation

Transformers

Transformers in vision

- Image is processed as a sequence of 16x16 patches
- Multi-head dot product self attention block can replace convolution

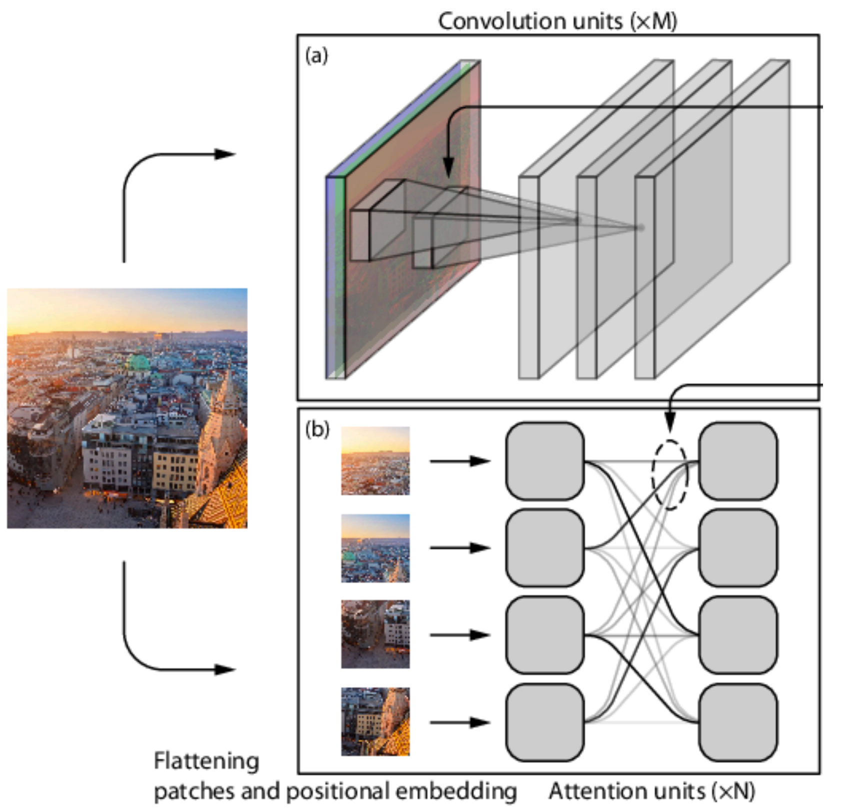




Transformers in vision

- Contrary to CNN, transformers are able to capture long-range dependencies
--> by computing attention score between any 2 patch representations

- They require more training data than CNN to generalize well: ViT trained on 300M images



Our idea: We want to analyze the information flow in transformer blocks
How can we use it to improve the design of the models and compress them?

Image sources: Tuli, Shikhar et al. “Are Convolutional Neural Networks or Transformers more like human vision?” *ArXiv* abs/2105.07197 (2021): n. pag.

Outline

- Presentation of transformer based segmentation models
- A bit deeper into attention: how can we visualize it?
- Presentation of the 3 datasets
- Results
  - Visualizing attention maps
  - Performance of compressed vs uncompressed models

(or dense prediction)

Transforming Transformers for image segmentation

- Principle: Remove the MLP head and use the transformer encoder layers

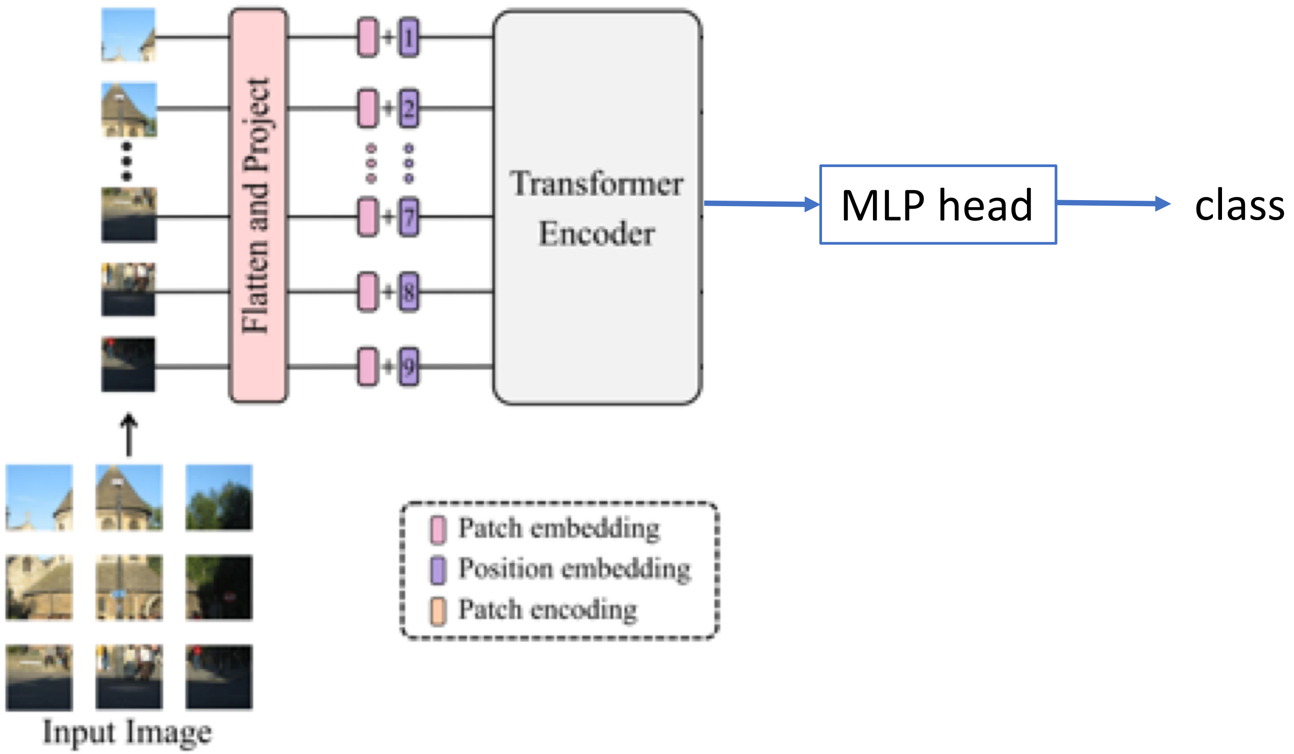


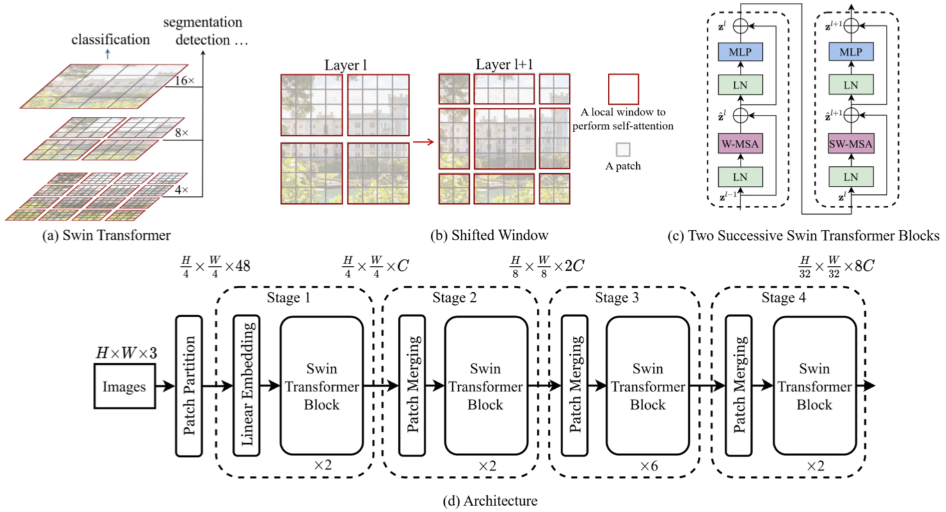
Image source: Strudel et al 2021 ICCV Segmenter: Transformer for Semantic Segmentation

Transforming Transformers for image segmentation

1) Pure transformer architecture

- Ex: **Swin Transformer**: Hierarchical Vision Transformer using Shifted Windows [Liu et al ICCV'21]

- Self-attention is computed within local windows.

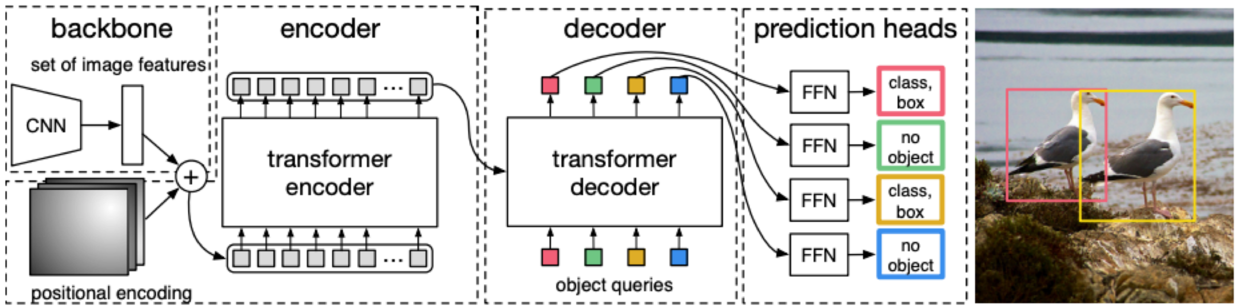


2) Hybrid: combine convolutional and transformer layers

Combine low-level CNN features + encodes strong global context

Ex: **DETR** End-to-End Object Detection with Transformers

[Carion et al ECCV '20]



Transformers in medical image segmentation

- Since 2021: Many papers proposing novel architectures based on pure transformers or hybrid CNN/transformers

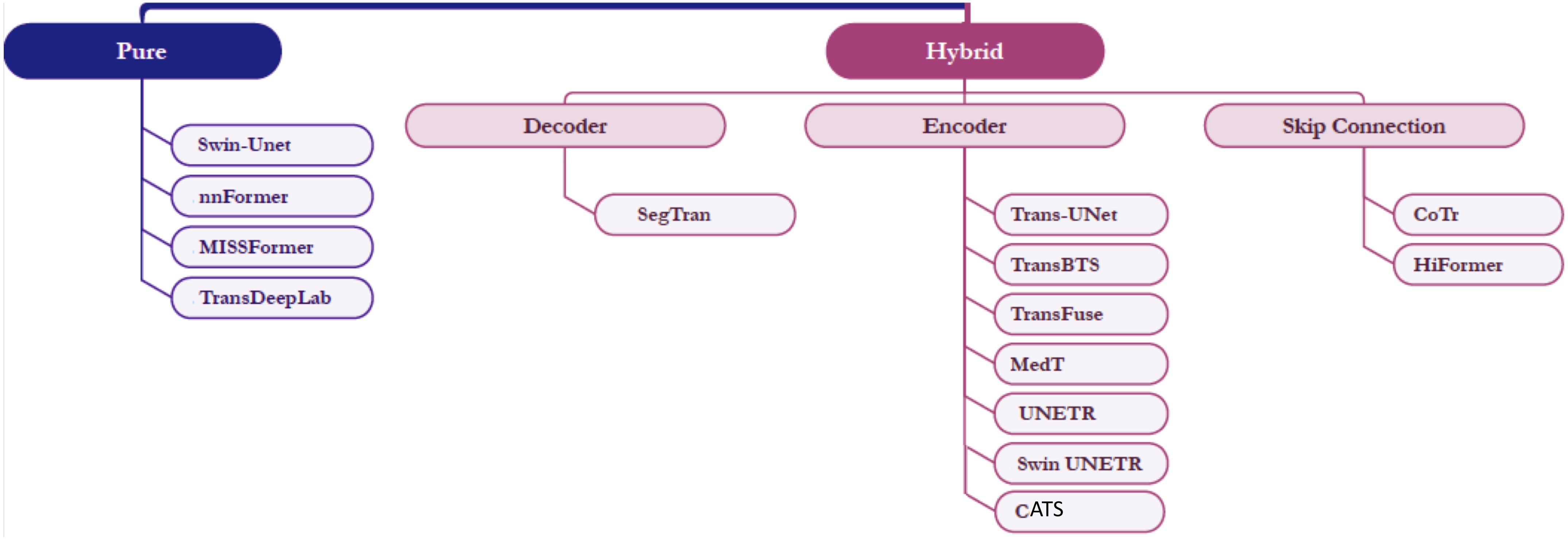
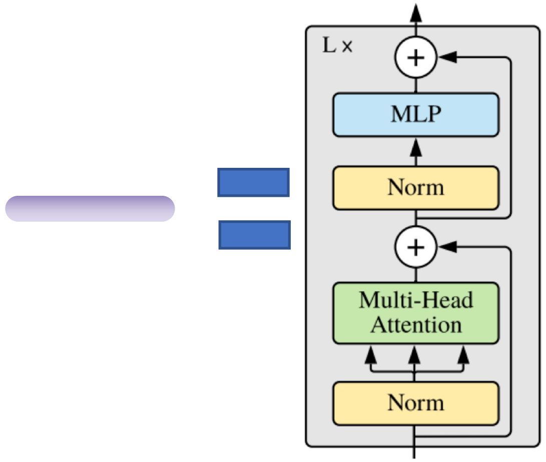


Figure modified from: Azad, Reza, et al. "Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review." *arXiv preprint arXiv:2301.03505* (2023).

Hybrid transformer + CNN



TransUNet

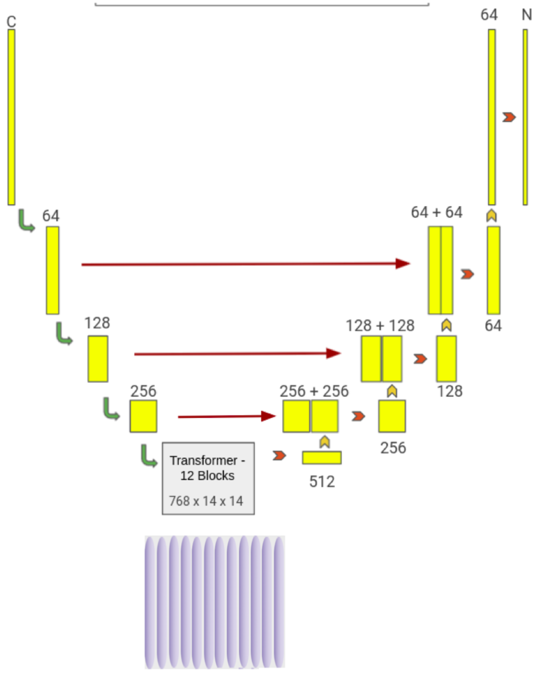
Chen et al 2021

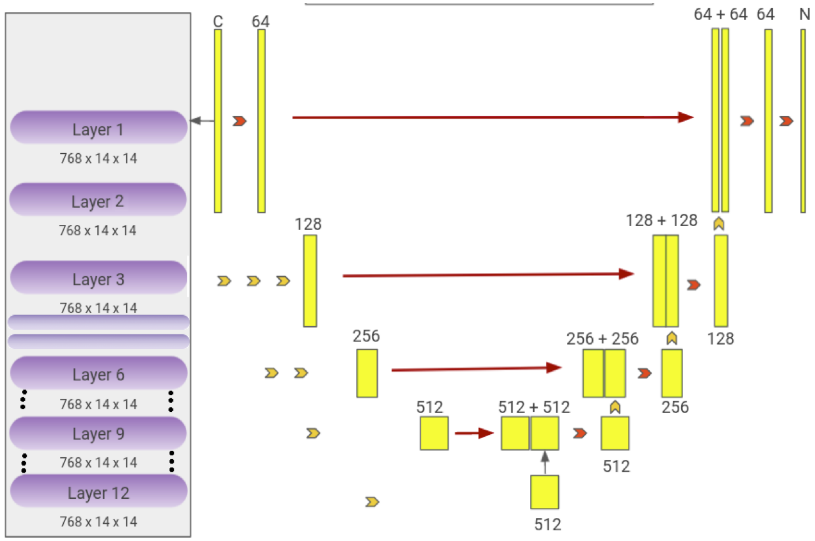
UNETR

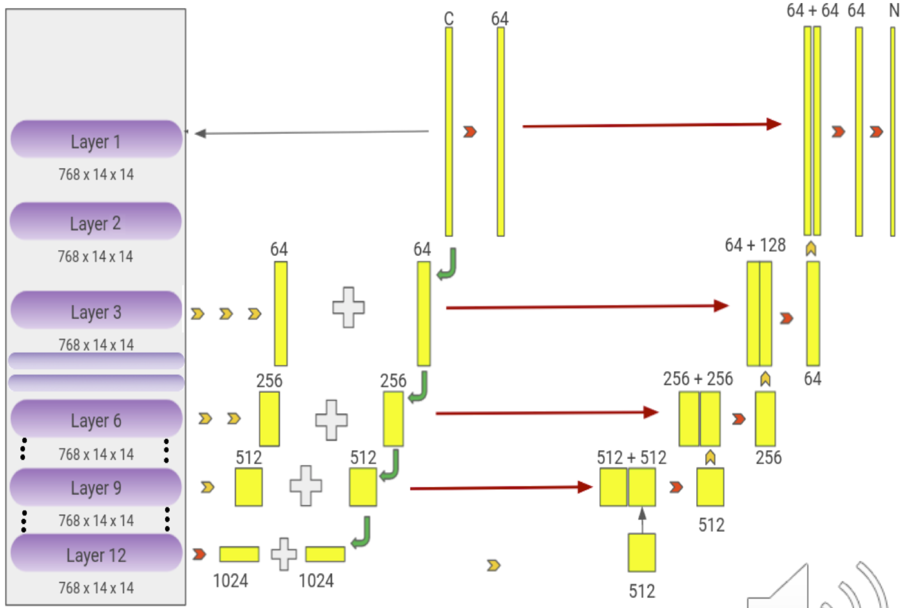
Li et al 2022

CATS

Hatamizadeh, et al 2022



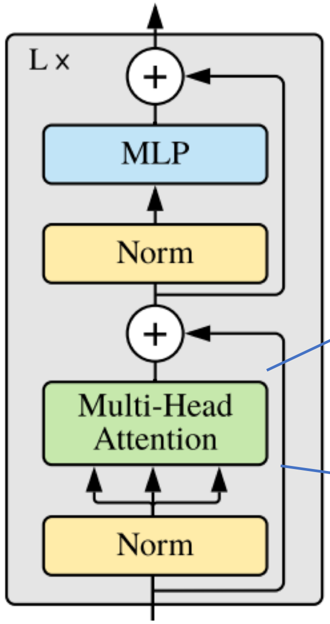


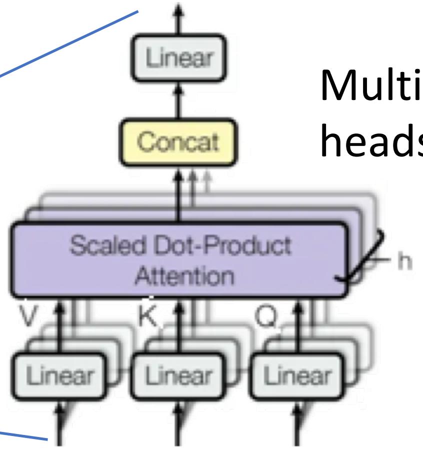


Deeper into transformer

« "attention distance" is analogous to receptive field size in CNNs.»
Dosovitskiy et al

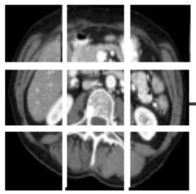
- Multi-head self attention





Multiple heads (h=8)

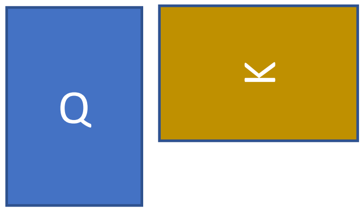
Patch embeddings + concat



Sequence z of N patches of size D

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

QK^T is called raw attention (dimension NxN)



Q, K, V are *D/h*-dimensional representation of the sequence

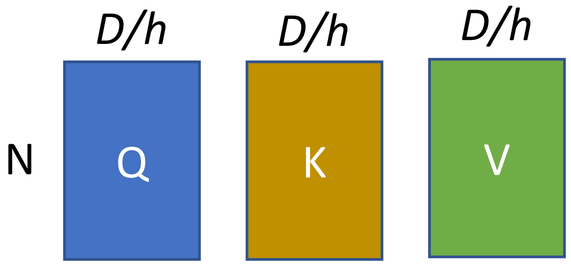


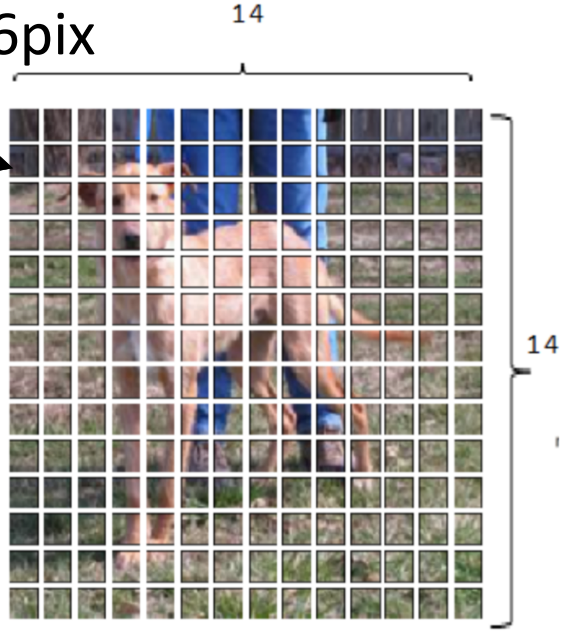
Image source: modified from TransUNet paper

Visualization of raw attention scores

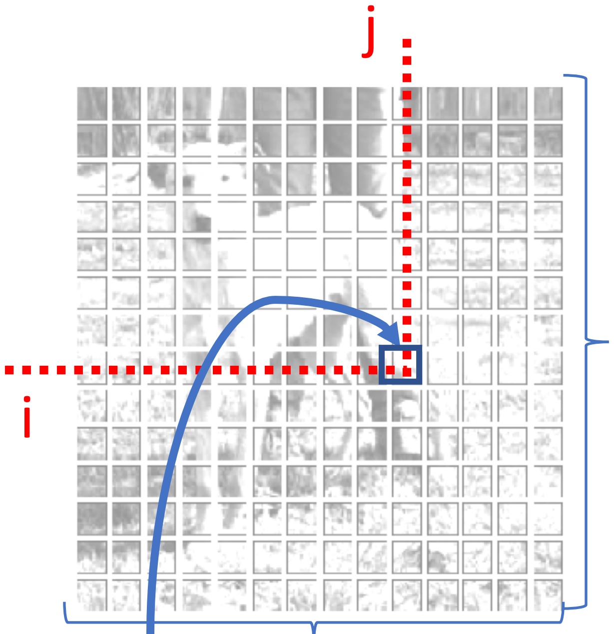
How does the receptive field evolve through the layers?

patch size: 16x16pix

224x224 image



patch nb: N=14x14

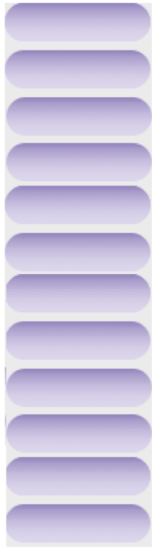


N=14x14 pix

N=14x14pix

Attention scores matrix for patch (i,j)

Stacking all L=12 layers:



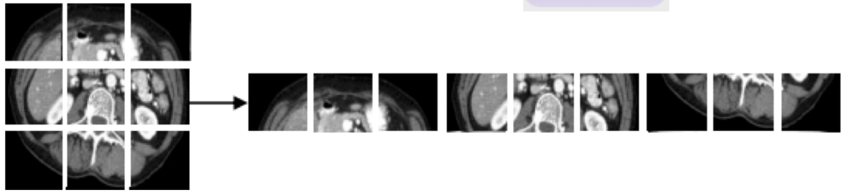
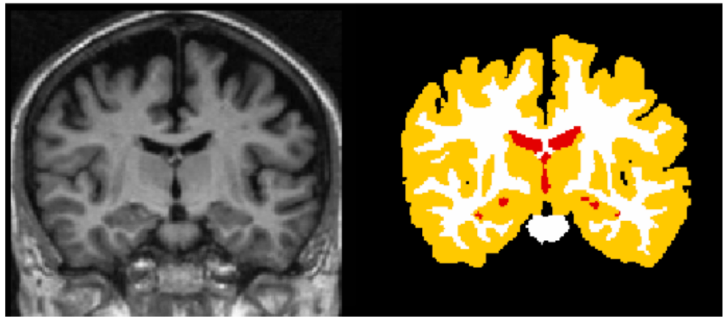


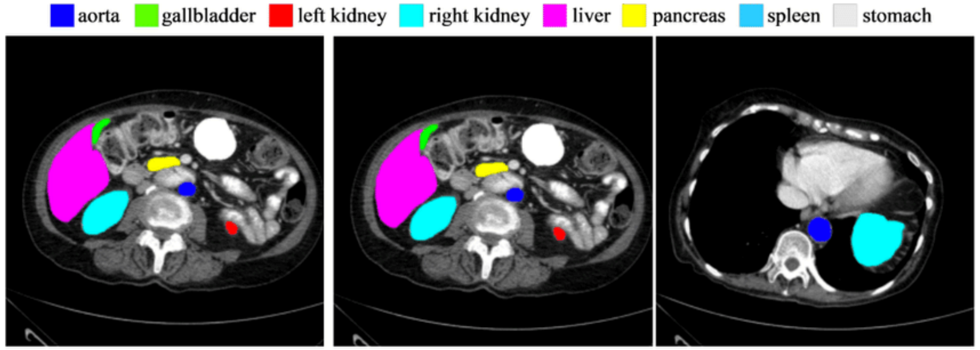
Image source: https://dkamatblog.home.blog/2021/08/05/vision-transformers-vit/

Datasets

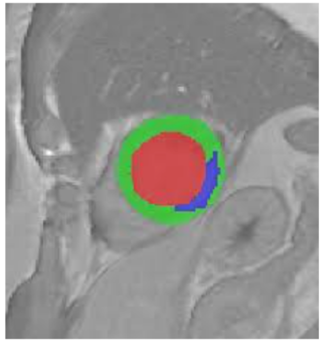
- IBSR 18: T1-weighted brain MRI images
  - 3 labels
  - 2D slices : 1280 train, 640 validation
- Synapse multi-organ: abdominal CT scans
  - 8 labels, 30 patients
  - 558 2D slices train, 180 validation.
- EMIDEC: delayed-enhancement cardiac MRI
  - 3 labels
  - 2211 2D slices train, 1568 validation

Cerebrospinal Fluid (CSF), Gray Matter (GM), and White Matter (WM)



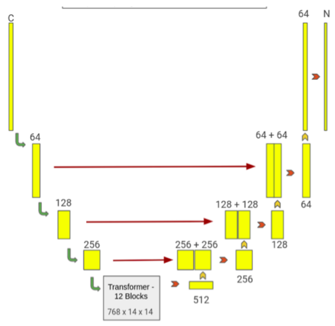


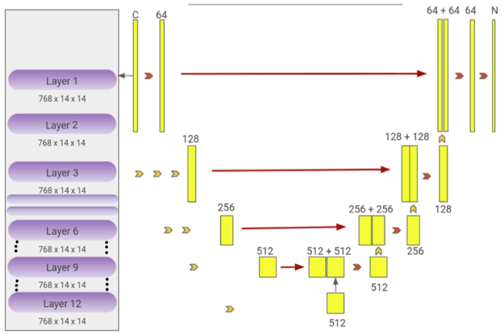
Myocardium, Infarction, and NoReflow



http://www.nitrc.org/projects/ibsr, https://www.synapse.org, http://emidec.com/dataset

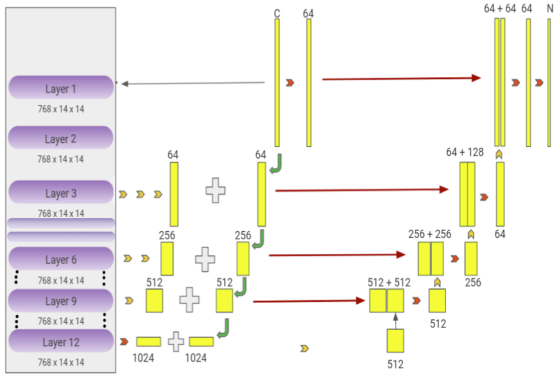
Experimental Protocol





- 3 hybdrid 2D models include

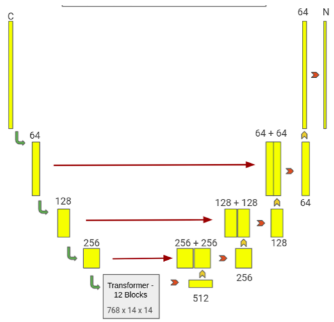
a pre-trained ViT with a ResNet backbone



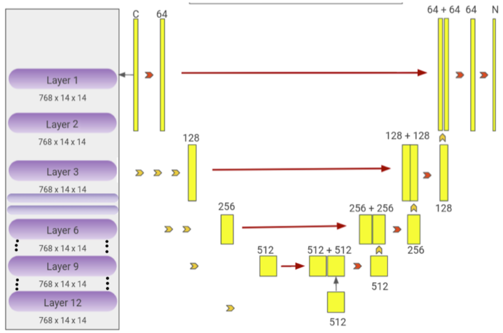
Configuration:

- AdamW optimizer, lr = 1e – 04
- Loss function: cross entropy + dice
- Epochs: 100 (IBSR 18, EMIDEC), 80 (Synapse multiorgan)
- Batchsize: 12

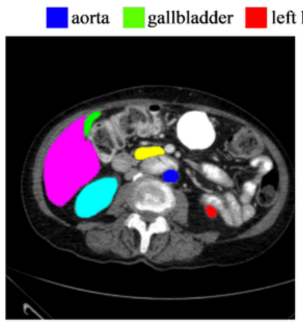
Attention score in Layer 1

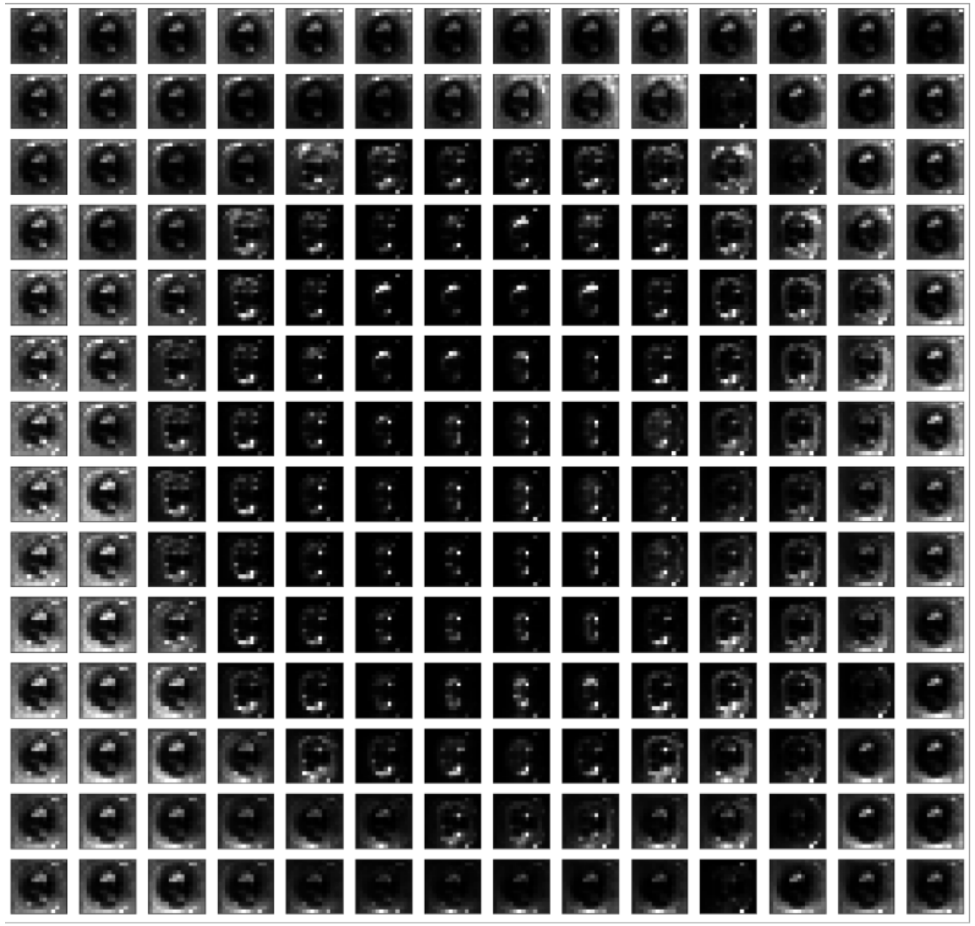


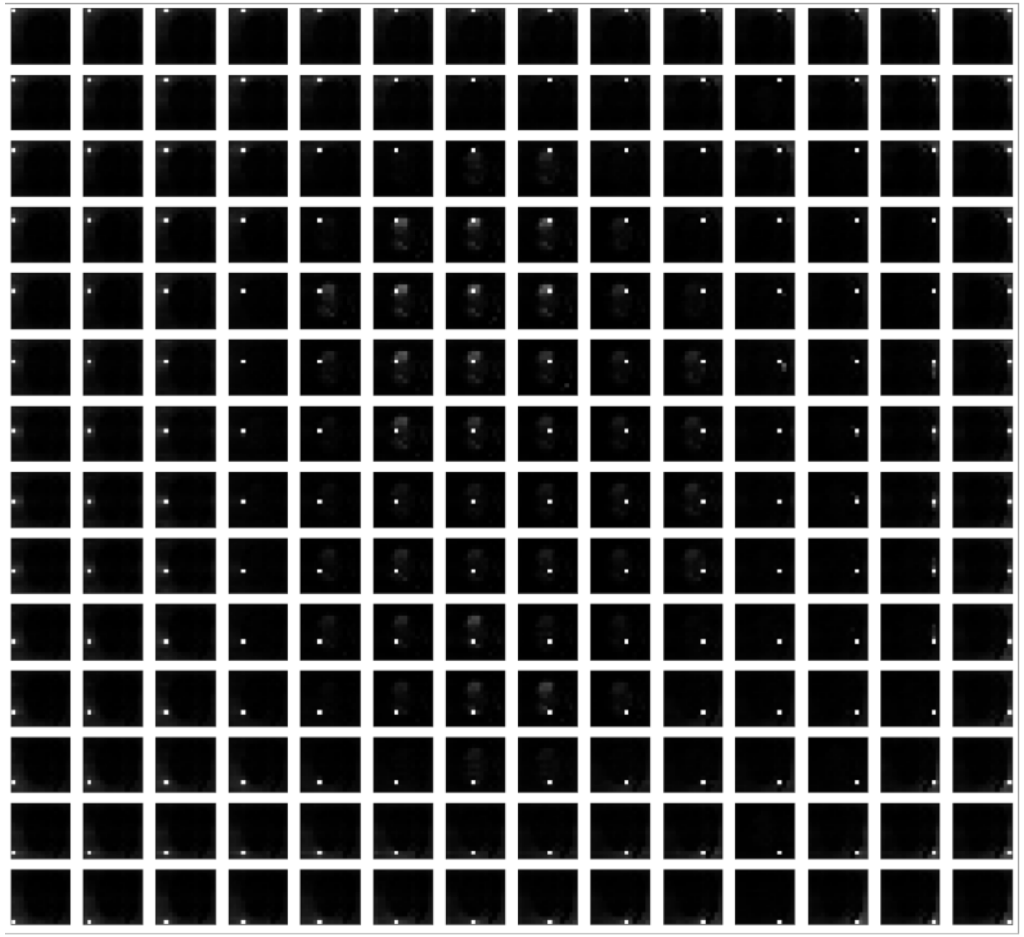
TransUNet

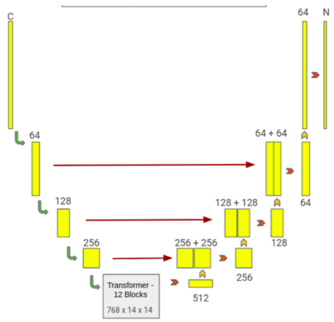


UNETR

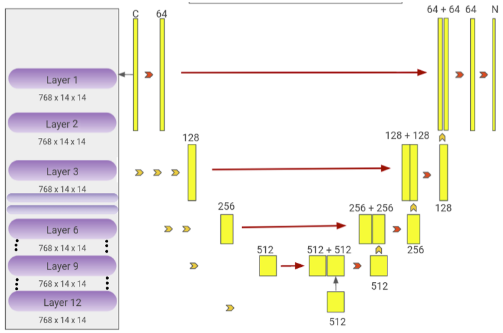






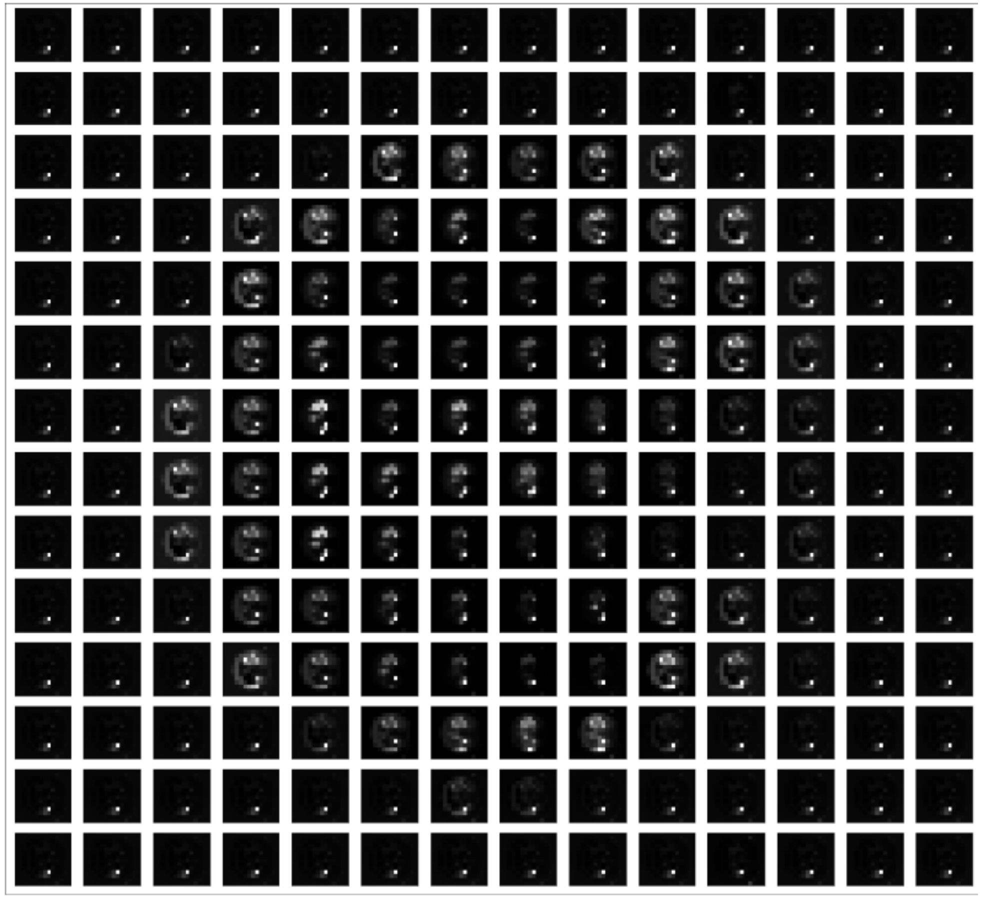


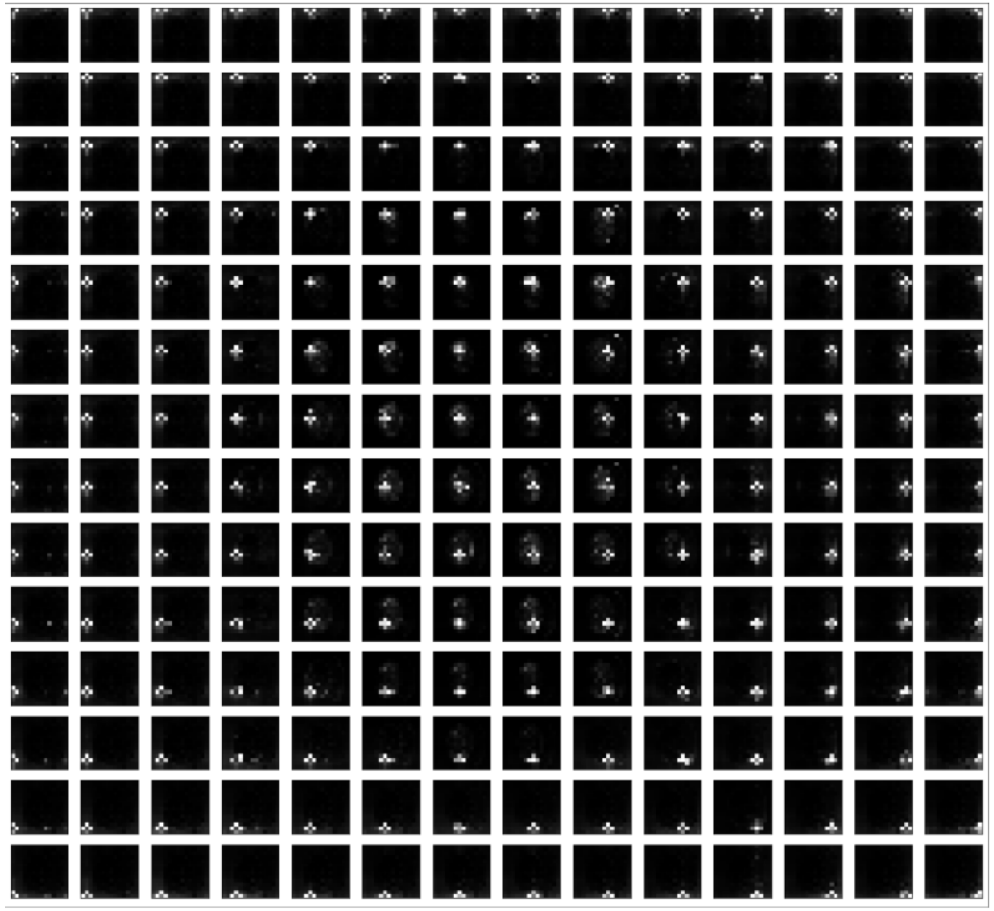
Attention score in Layer 2

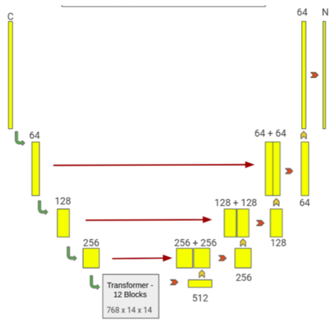


TransUNet

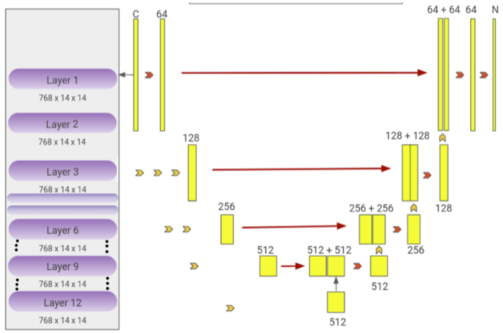
UNETR



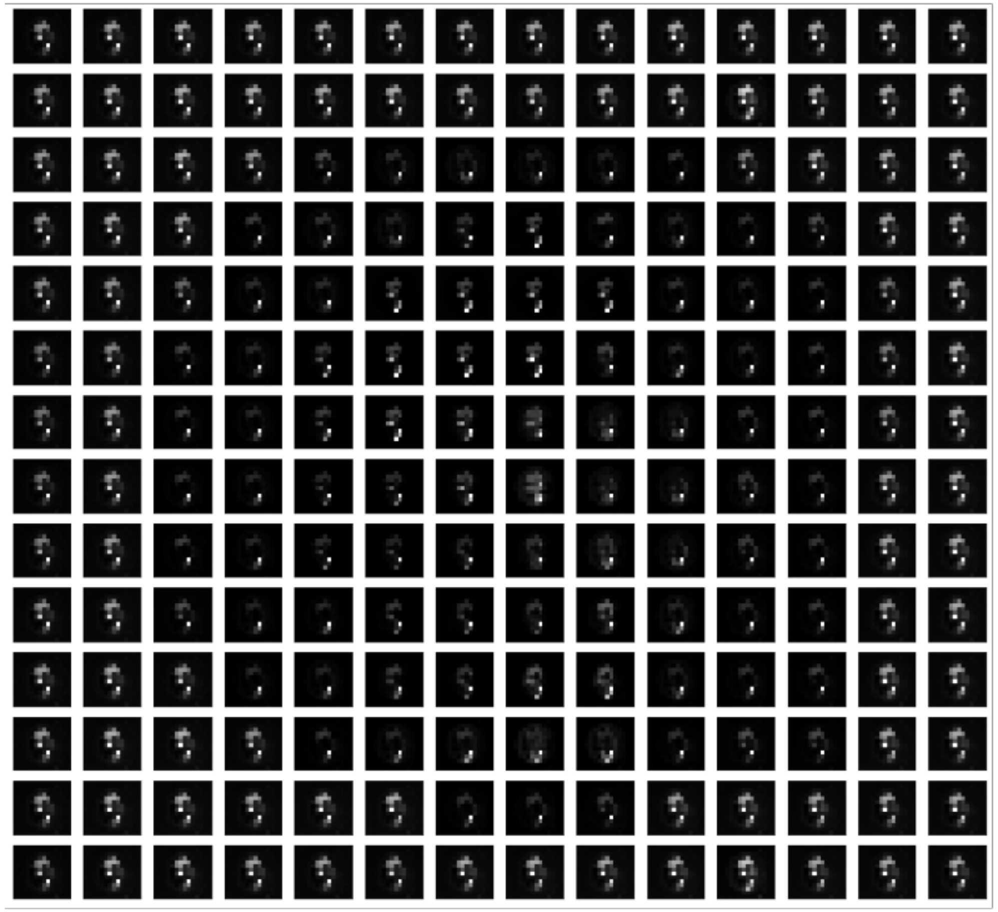




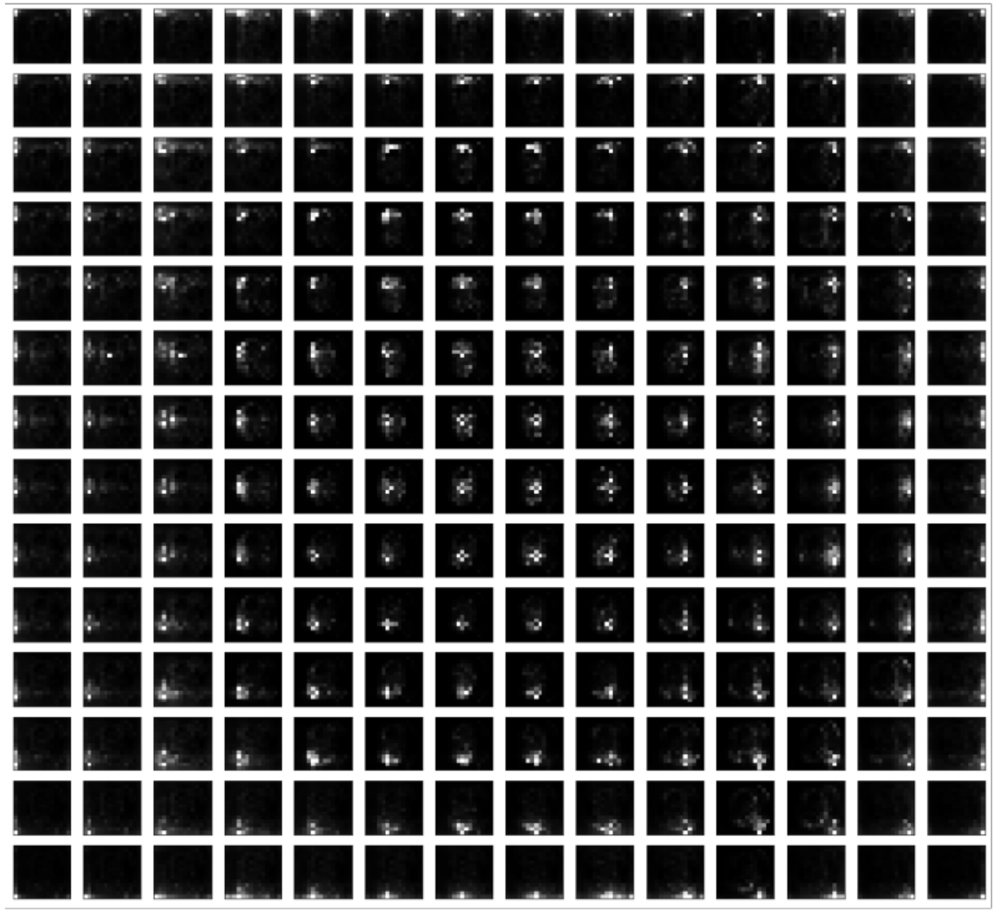
Attention score in Layer 3

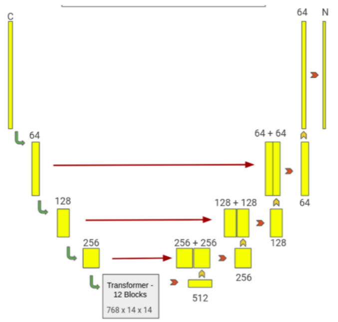


TransUNet

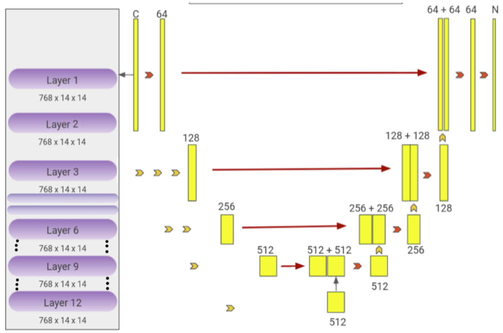


UNETR



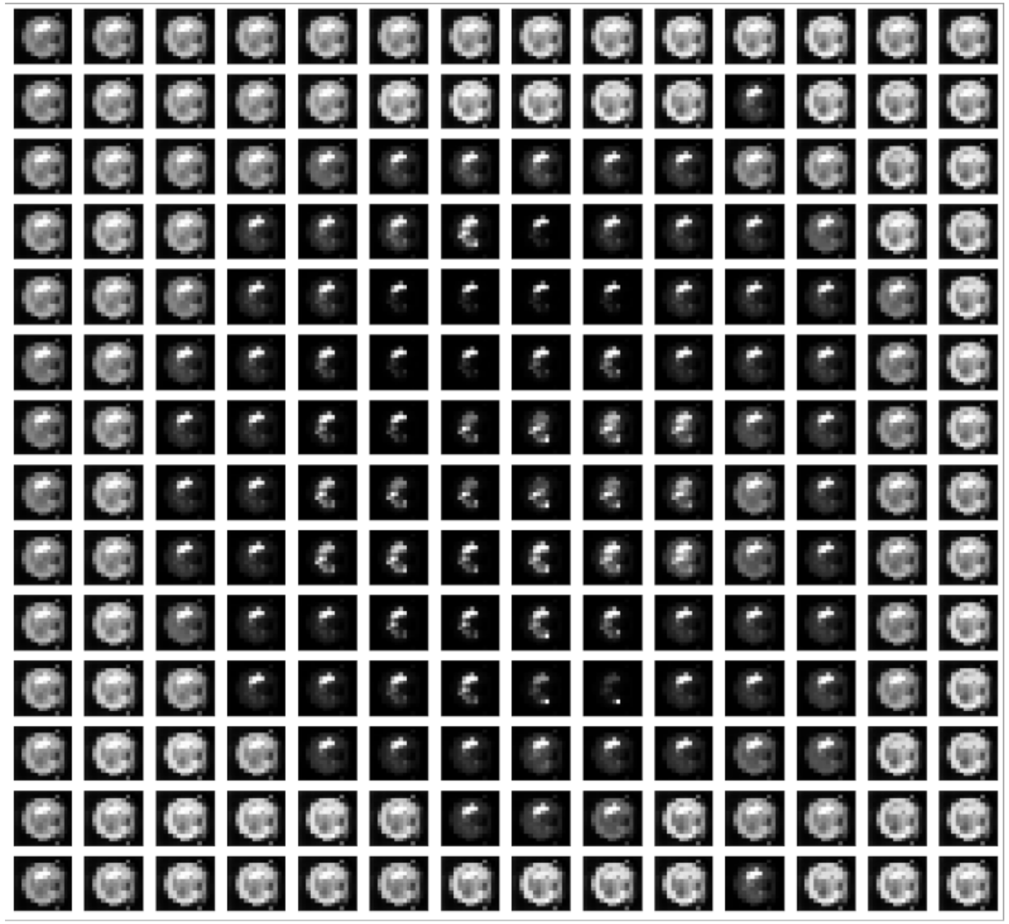


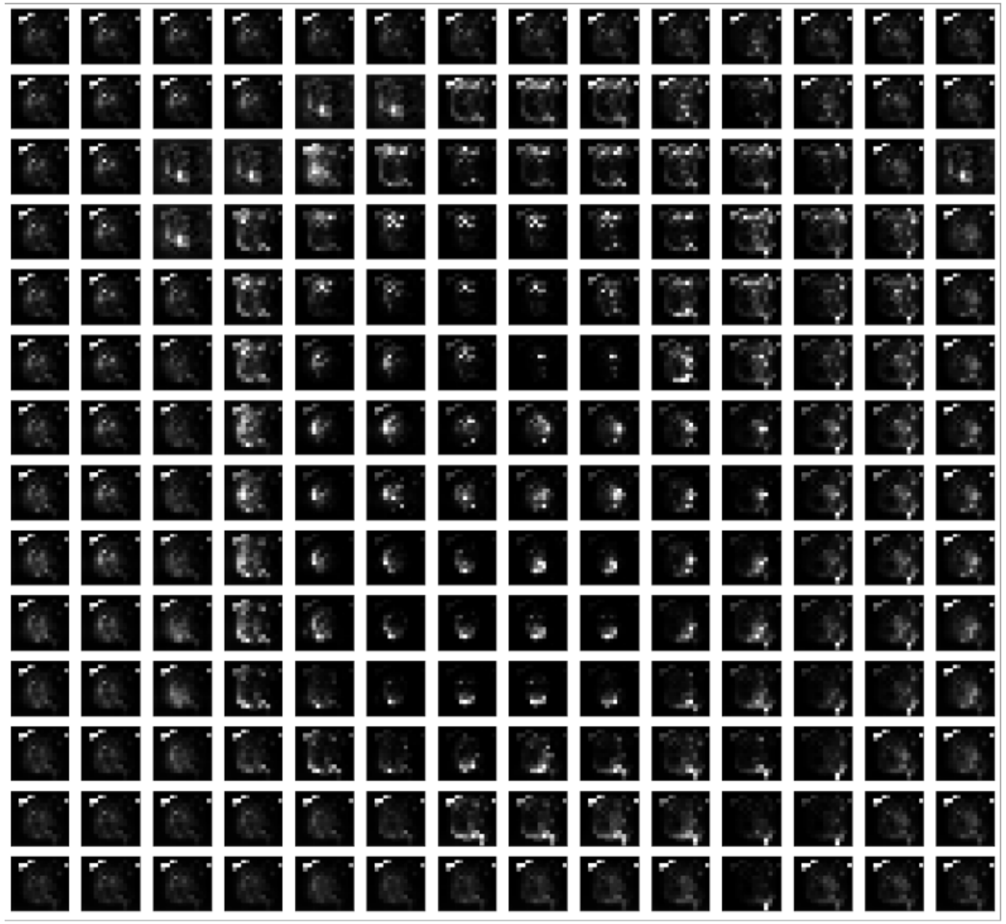
Attention score in Layer 6



TransUNet

UNETR





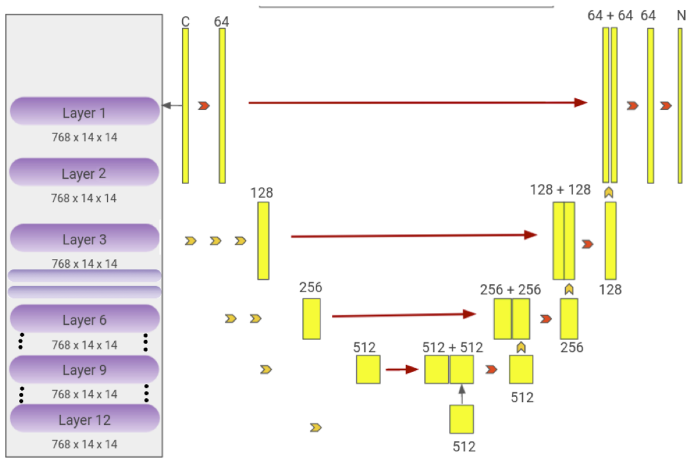
Proposition

- TransUNet: receptive field seems global, starting from the first block!

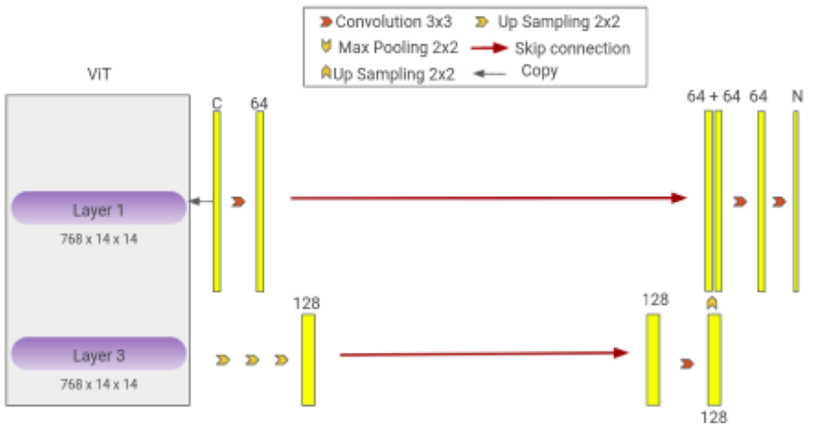
→ Keep 1 and 3 blocks

- UNETR receptive field starts to be global after 3 blocks

→ Keep only 3 blocks instead of 12



Uncompressed UNETR



Compressed UNETR

Results: Average Dice on the organs

Model	#B	IBSR 18	%D	%PS	EMIDEC	%DC	%PS	Synapse	%DC	%PS
TransUNet	12	0,860	-	-	0,758	-	-	0,818	-	-
2D UNETR	12	0,836	-	-	0,719	-	-	0,768	-	-
2D CATS	12	0,862	-	-	0,699	-	-	0,810	-	-

Results on compressed models

Change % in the number of parameters

Model	#B	IBSR 18	%DC	%PS	EMIDEC	%DC	%PS	Synapse	%DC	%PS
TransUNet	12	0,860	_	_	0,758	_	_	0,818	_	_
TransUNet	3	0,860	0	↓61	0,760	↑+ 0,2	↓61	0,824	↑+0,7	↓61
2D UNETR	12	0,866	_	_	0,719	_	_	0,768	_	_
2D UNETR	3	0,864	↓-0.2	↓75	0,703	↓-2,2	↓75	0,786	↓-1,7	↓75
2D CATS	12	0,862	_	_	0,699	_	_	0,810	_	_
2D CATS	3	0,864	↑+ 0.2	↓61	0,720	↑+2.9	↓65	0,788	↓-2,6	↓63

Results on compressed models

Change % in the number of parameters

Model	#B	IBSR 18	%DC	%PS	EMIDEC	%DC	%PS	Synapse	%DC	%PS
TransUNet	12	0,860	–	–	0,758	–	–	0,818	–	–
TransUNet	3	0,860	0	↓61	0,760	↑+ 0,2	↓61	0,824	↑+0,7	↓61
2D UNETR	12	0,866	–	–	0,719	–	–	0,768	–	–
2D UNETR	3	0,864	↓-0.2	↓75	0,703	↓-2,2	↓75	0,786	↓-1,7	↓75
2D CATS	12	0,862	–	–	0,699	–	–	0,810	–	–
2D CATS	3	0,864	↑+ 0.2	↓61	0,720	↑+2.9	↓65	0,788	↓-2,6	↓63

Results on compressed models

Keeping only one transformer block in TransUNet allows to reduce the nb of parameters by 74%

Model	#B	IBSR 18	%DC	%PS	EMIDEC	%DC	%PS	Synapse	%DC	%PS
TransUNet	12	0,860	-	-	0,758	-	-	0,818	-	-
TransUNet	3	0,860	0	↓61	0,760	↑+ 0,2	↓61	0,824	↑+0,7	↓61
TransUNet	1	0,865	↑+ 0,6	↓74	0,769	↑+1,4	↓74	0,824	↑+0,7	↓74
2D UNETR	12	0,836	-	-	0,719	-	-	0,768	-	-
2D UNETR	3	0,864	↓-0.2	↓75	0,703	↓-2,2	↓75	0,786	↓-1,7	↓75
2D CATS	12	0,862	-	-	0,699	-	-	0,810	-	-
2D CATS	3	0,864	↑+ 0.2	↓61	0,720	↑+2.9	↓65	0,788	↓-2,6	↓63

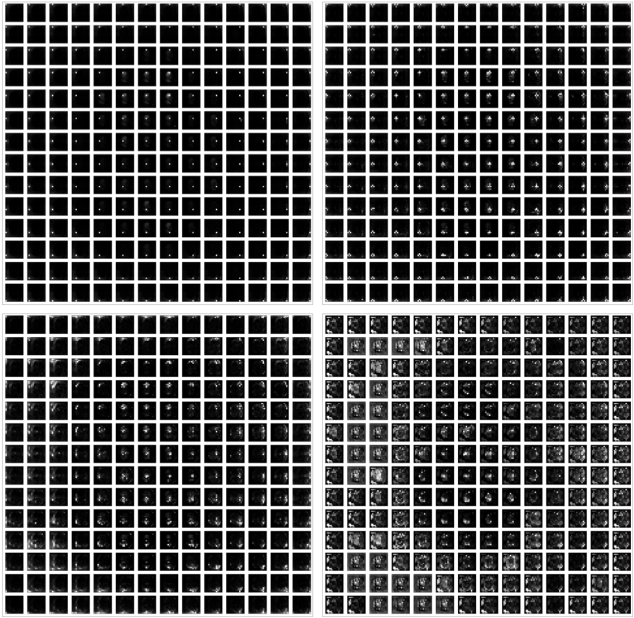
Conclusion

- Attention information from transformer blocks is helpful
  - towards analyzing information flow
  - to compress the model without seriously sacrificing model performance

- Not necessary to have all 12 transformer blocks in order to achieve a global receptive field

  - Compressed versions have < 50% of the original parameters.

Perspectives



- Limitation: Qualitative analysis of the receptive field

- Explainability for transformer-based segmentation models
- Visualizing attention scores (inner products of queries and keys) reduces greatly the information
  - ‘attention rollout’: summarises the various attention maps throughout the layers.[Abnar ACL 2020]
  - Consider also other layers [Chefer CVPR 2021]

Thank you for your attention!

This is a joint work with:

Syed Nouman Hasany

Fabrice Mériaudeau

Results are published in:

Hasany, S. N., Petitjean, C., & Meriaudeau, F. (2023). A study of attention information from transformer layers in hybrid medical image segmentation networks. In *SPIE Medical Imaging, San Diego*.

- The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant Project-ANR-21-CE23-0013 (project MediSEG).

Caroline.Petitjean@univ-rouen.fr