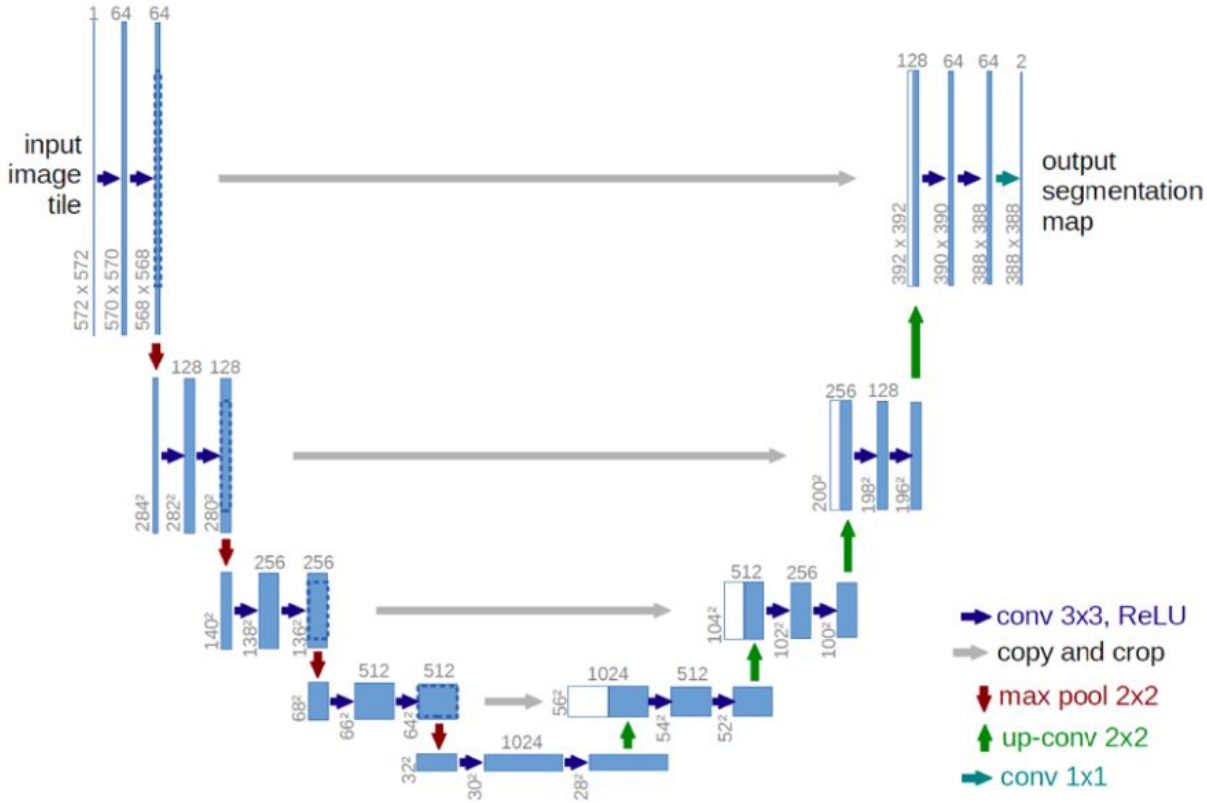
TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation

Authors: Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou.

Paper submitted on 08.02.2021

The presentation was prepared by: Sergey Pnev

U-net


input image tile
output segmentation map
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

CNN(U-net) problems

- CNN-based approaches generally exhibit limitations for modeling explicit long-range relation, due to the intrinsic locality of convolution operations.
- Weak performances especially for target structures that show large inter-patient variation in terms of texture, shape and size.

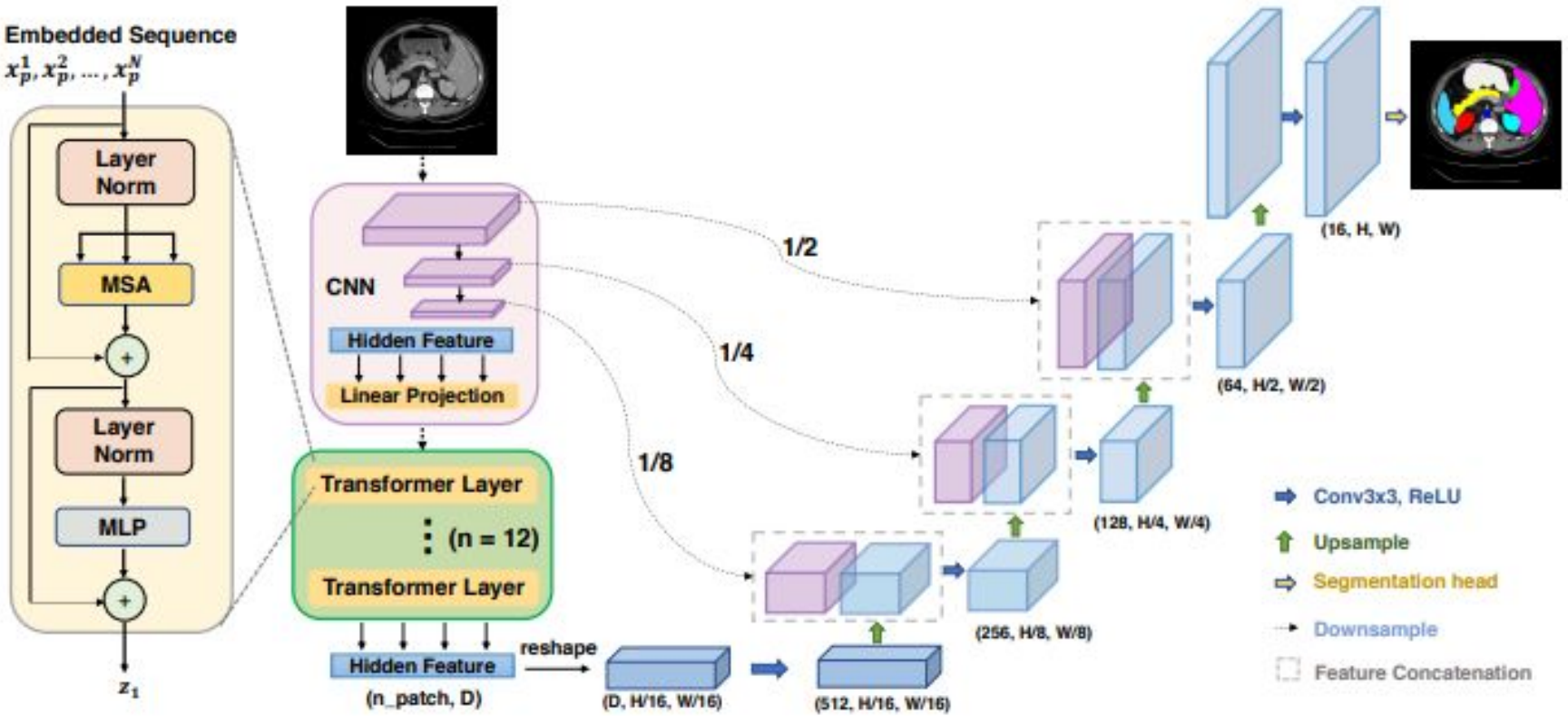
Solution

- Transformers!
- They are powerful at modeling global contexts.
- And demonstrate superior transferability for downstream tasks under large-scale pre-training.

But...

- Transformers treat the input as 1D sequences and exclusively focus on modeling the global context at all stages
- Not enough localization details in low-resolution features.
- And this information cannot be effectively recovered by direct upsampling to the full resolution
- This leads to a coarse segmentation outcome.
- So, naive usage (i.e., use a transformer for encoding the tokenized image patches, and then directly upsamples the hidden feature representations into a dense output of full resolution) cannot produce a satisfactory result.

U-net + Transformer = TransUNet



Math

- Input $x \in R^{H \times W \times C}$
- Output $y \in R^{H \times W \times C'}$
- Reshaping the input x into a sequence of flattened 2D patches

$\{x_p^i \in R^{P^2 \times C} | i = 1, \ldots, N\}$, where each patch is of size $P \times P$ and $N = \frac{HW}{P^2}$ is the number of image patches

Positional embeddings

- Map vectorized patches x_p into a latent D-dim embedding space using a trainable linear projection. To encode the patch spatial information, we learn specific position embeddings which are added to the patch embedding to retain positional information as follows:

$$z_0 = \left[x_p^1 E; \ldots; x_p^N E\right] + E_{pos}$$

- $E \in R^{(P^2 C) \times D}$ is the patch embedding projection, and $E_{pos} \in R^{N \times D}$ denotes the position embedding.

Transformer encoder

- Consists of L layers of Multihead Self-Attention(MSA) and Multi-Layer Perceptron blocks.
- The output of l-th layer:

$$z_l' = MSA\left(LN(z_{l-1})\right) + z_{l-1}$$
$$z_l = MLP\left(LN(z_l')\right) + z_l'$$

MSA

- Input: $C = z_0 = \left[x_p^1 E; \ldots; x_p^N E\right] + E_{pos} = [c_1, \ldots, c_N]$
- Let $C \in R^{N \times D}$
- Matrices: $W_q \in R^{D \times K}, W_k \in R^{D \times K}, W_v \in R^{D \times D}$

$$Q = CW_q$$
$$K = CW_k$$
$$V = CW_v$$
$$AttentionMap = A = \sigma(QK^T)$$
$$Output = AV$$

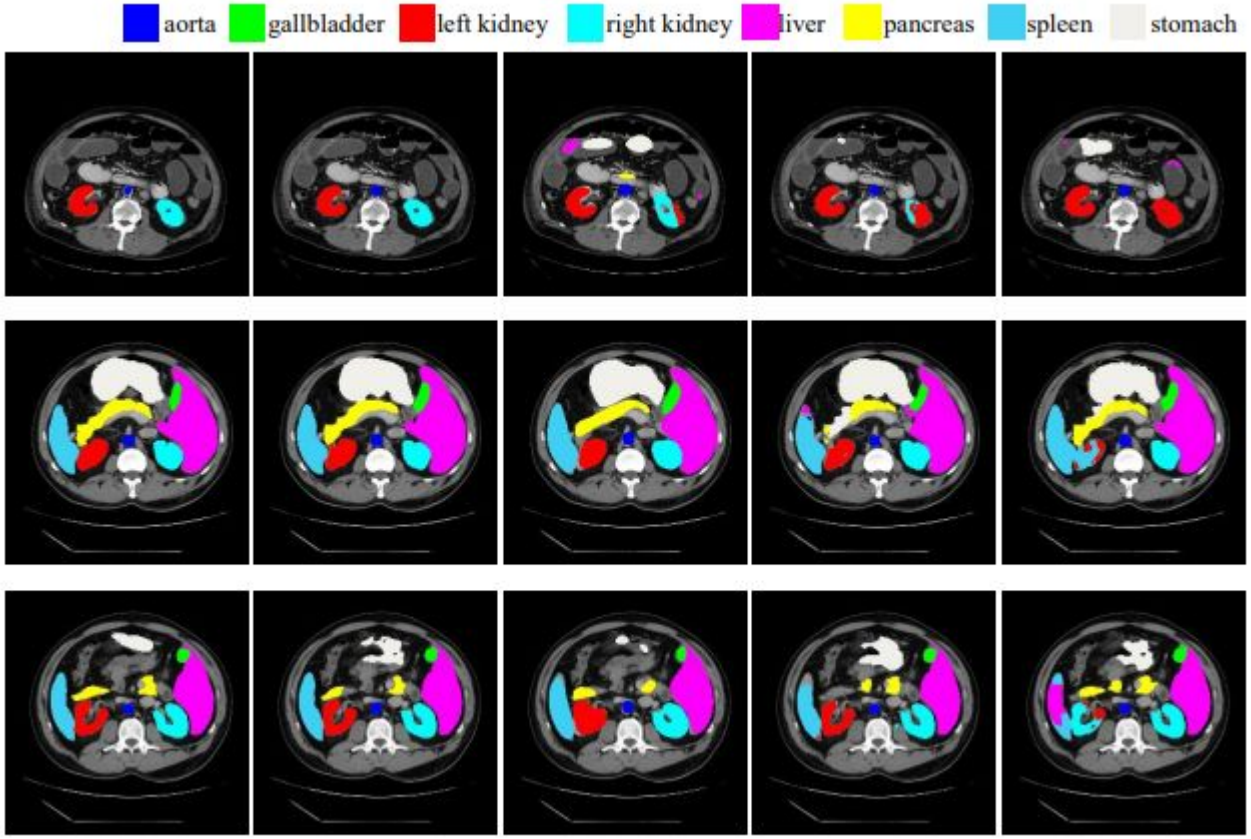
Implementation details

- Synapse multi-organ segmentation dataset, CT, 30 scans, 18 - training, 12 - validation. About 2212 slices.
- Automated cardiac diagnosis challenge, MRI, 70 - training, 10 - validation, 20 - test. 1930 axis slices for training.
- Augmentation: random rotation, flipping
- All models and transformers were pretrained on ImageNet
- SGD optimizer with l_r = 0.001, momentum 0.9 and weight decay 1e-4.
- Batch size = 24, 20k epochs for ACDC and 14k for Synapce.

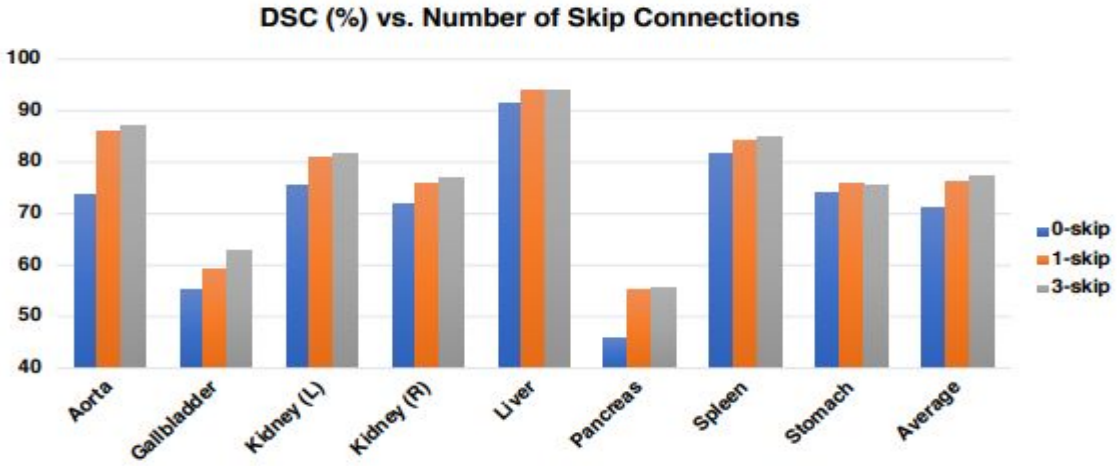
Results

Framework		**Average**		Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Encoder	Decoder	DSC ↑	HD ↓								
V-Net [9]		68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [5]		69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50	U-Net [12]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50	AttnUNet [13]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
ViT [4]	None	61.50	39.61	44.38	39.59	67.46	62.94	89.21	43.14	75.45	69.78
ViT [4]	CUP	67.86	36.11	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50-ViT [4]	CUP	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet		**77.48**	**31.69**	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62

Results


aorta
gallbladder
left kidney
right kidney
liver
pancreas
spleen
stomach
(a) GroundTruth
(b) TransUNet
(c) R50-ViT-CUP
(d) AttnUNet
(e) UNet

Analytical study



- ACDC dataset

Framework	Average	RV	Myo	LV
R50-U-Net	87.55	87.10	80.63	94.92
R50-AttnUNet	86.75	87.58	79.20	93.47
ViT-CUP	81.45	81.46	70.71	92.18
R50-ViT-CUP	87.57	86.07	81.88	94.75
TransUNet	89.71	88.86	84.53	95.73

Analytical study

- Image resolution

Resolution	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
224	**77.48**	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
512	**84.36**	90.68	71.99	86.04	83.71	95.54	73.96	88.80	84.20

- Patch size and seq. length

Patch size	Seq_length	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
32	49	76.99	86.66	63.06	81.61	79.18	94.21	51.66	85.38	74.17
16	196	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
8	784	**77.83**	86.92	58.31	81.51	76.40	93.81	58.09	87.92	79.68

- Model scale

Model scale	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Base	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Large	78.52	87.42	63.92	82.17	80.19	94.47	57.64	87.42	74.90

Conclusion

The authors of the paper presented one of the first works on the application of transformers to the medical images semantic segmentation task and showed successful results on 2 datasets

Thank you!