MEDICAL TRANSFORMERS

Ahmet Emir Kocaaga, Ramiz Dundar[1]

[1]Advisor: Pinar Yanardag, Fatma Basak Aydemir

Introduction

- Transformers [3] are a widely adopted technique in the natural language processing research community due to their powerful attention and parallelization capabilities.
- In the computer vision scene, before vision transformer (ViT) [1], pure applications of transformers on images often are underperformed compared to convolutional neural networks or hybrid models.
- Vision transformer (ViT), although a relatively new approach, has the state of the art results in image classification.
- We apply vision transformers (ViT) to the medical domain for the first time.
- We also propose a strided (SViT) approach that improves both vision transformer (ViT) and vision transformer trained with distillation tokens (DViT) [2].

Dataset

- We use the Irvin et al. (2019) dataset. Chexpert is a large chest X-ray imagedataset with 224,316 chest radiographs of 65,240 patients.
- Each X-ray image is either takenfrom the side (lateral) or taken from the front (frontal).

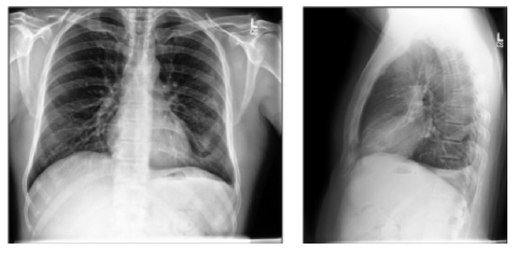

Fig. 1: Chexpert dahaset example image.

- The National Institutes of Health's Clinical Center CT image dataset.
- DeepLesion dataset has over 32,000 publicly available CT images.

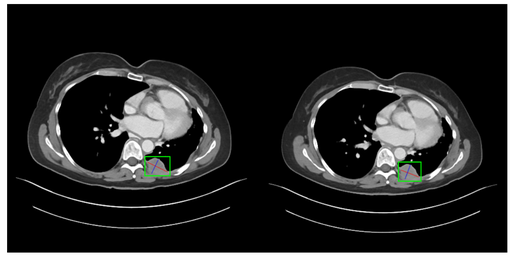

Fig. 2: Deeplesion dahaset example image.

Methodology

- Dosovitskiy et al. [1] proposed a new approach for transformers in visual domain.
- Instead of receiving a 1D sequence of tokens, their patch-based strategy reshapes the image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\mathbf{x_p} \in \mathbb{R}^{N \times P^2 C}$ where P where (H, W) represents the size of the original image, (P, P) represents the size of each image patch, and $N = HW/P^2$ represents the length of the sequence for the Transformer.

$$\mathbf{z}_0 = [\mathbf{x}_{class}; \mathbf{x}_p^1\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E};] + \mathbf{E}_{pos} \quad (1)$$

$$\mathbf{z}'_l = \mathbf{MSA}(\mathbf{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1} \quad (2)$$

$$\mathbf{z}_l = \mathbf{MLP}(\mathbf{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l \quad (3)$$

$$\mathbf{y} = \mathbf{LN}(\mathbf{z}_l^0) \quad (4)$$

- $\mathbf{MSA}$ represents multi-headed self attention, $\mathbf{MLP}$ represents multi-layer percepetron that contains two layers with a GELU non-linearity. A learnable embedding $\mathbf{z}_0^0$ is appended to the sequence of embedded patches and $\mathbf{z}_0^L$ represents the image representation denoted with $\mathbf{y}$. $\mathbf{MSA}$ is an extension of self-attention that projects the outputs of multiple self-attention operations run in parallel.

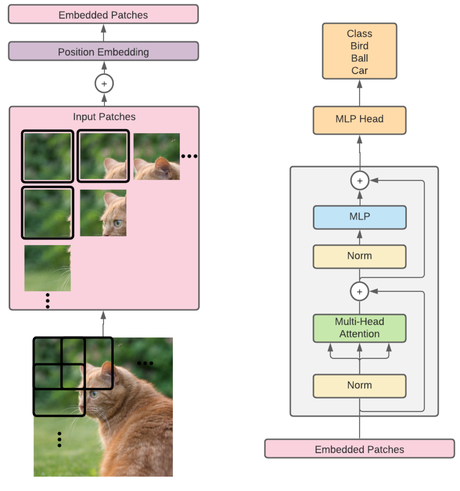

Fig. 3: Strided Vision Transformer.

- We propose a "strided" approach to provide memory efficiency
- In this approach, we again split the image into the patches, but this time those parts of those patches overlap, in a strided fashion.

Working on 3D Images

- To use our classification model on 3D CT images, we need to convert 3D images to 2D vectors. We tried following approaches.
- Calculate average of the 2D image vectors in 3D CT scans.
- Converting 3D to 2D by concatenating slices.
- Using 3D to 2D autoencoder.

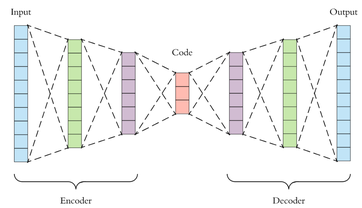

Fig. 4: 3D to 2D autoencoder.

Results

- We tried Vision Transformer (ViT) and our approach, Strided ViT.
- We further improved the results with Distillable ViT and Disttillable Strided ViT.

	Cardiomegaly	Edema	Consolidation	Atelectasis	Pleural Effusion	Mean AUC
ViT	0.825	0.873	0.888	0.815	0.888	0.858
S-ViT	0.809	0.899	0.900	0.828	0.898	0.867
D-ViT	0.827	0.892	0.896	0.824	0.909	0.870
DS-ViT	0.829	0.894	0.922	0.836	0.917	0.880

Fig. 5: 2D Classification results.

	Mean AUC
Vector Mean	0.437
Concatenation	0.611
3D to 2D	0.727

Fig. 6: 3D Classification results.

References

[1] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[2] Hugo Touvron et al. "Training data-efficient image transformers & distillation through attention". In: (2021). arXiv: 2012.12877 [cs.CV].

[3] Ashish Vaswani et al. "Attention is all you need". In: (2017), pp. 5998–6008.