**Is it Time to Replace CNNs with Transformers for Medical Images?**

* The transformer neural network is a novel architecture that aims to solve sequence-to- sequence tasks while handling long-range dependencies with ease.
* Recently, vision transformers (ViTs) have appeared as a competitive alternative to CNNs, yielding similar levels of performance while possessing several interesting properties that could prove beneficial for medical imaging tasks.
* The attention mechanism central to transformers offers several key advantages over convolutions:

(1) it captures long-range relationships,

(2) it has the capacity for adaptive modelling via dynamically computed self-attention weights that capture relationships between tokens,

(3) it provides a type of built-in saliency which gives insight as to what the model focused on.

* Yet, evidence suggests that vision transformers require very large datasets to outperform CNNs – in, the benefits of ViT only became evident when Google’s private 300 million image dataset, JFT-300M, was used for pretraining. Their reliance on data of this scale is a barrier to the widespread application of transformers.
* CNNs pre-trained on IMAGENET typically outperform those trained from scratch in the medical domain, both in terms of final performance and reduced training time.
* As expected, ViTs are worse than CNNs in the low data regime if one simply trains from scratch.
* Transfer learning bridges the performance gap between CNNs and ViTs; performance is similar.
* The best performance is obtained with self-supervised pre-training + fine-tuning, where ViTs enjoy a small advantage over comparable CNNs.
* More precisely, ViTs can reach the same level of performance as CNNs in small medical datasets, but require transfer learning in order to do so.
* However, using ImageNet pretrained weights is the standard approach for CNNs as well, so the switch to ViTs is trivial.
* Furthermore, ViTs can outperform CNNs using SSL pre-training when working with limited number of samples, but only marginally.