Radiology: Artificial Intelligence

SPECIAL REPORT

Toward Foundational Deep Learning Models for Medical Imaging in the New Era of Transformer Networks

*Martin J. Willemink, MD, PhD** • *Holger R. Roth, PhD** • *Veit Sandfort, MD*

From the Department of Radiology, Center for Academic Medicine, Stanford University School of Medicine, 453 Quarry Rd, 324-7, Palo Alto, CA 94304 (M.J.W., V.S.); Segmed, Menlo Park, Calif (M.J.W.); and NVIDIA, Santa Clara, Calif (H.R.R.). Received November 10, 2021; revision requested January 6, 2022; revision received October 16; accepted October 18. **Address correspondence to** V.S. (email: *veit.sandfort@gmail.com*).

*M.J.W. and H.R.R. contributed equally to this work.

Authors declared no funding for this work.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2022; 4(6):e210284 • https://doi.org/10.1148/ryai.210284 • Content code: AI

Deep learning models are currently the cornerstone of artificial intelligence in medical imaging. While progress is still being made, the generic technological core of convolutional neural networks (CNNs) has had only modest innovations over the last several years, if at all. There is thus a need for improvement. More recently, transformer networks have emerged that replace convolutions with a complex attention mechanism, and they have already matched or exceeded the performance of CNNs in many tasks. Transformers need very large amounts of training data, even more than CNNs, but obtaining well-curated labeled data is expensive and difficult. A possible solution to this issue would be transfer learning with pretraining on a self-supervised task using very large amounts of unlabeled medical data. This pretrained network could then be fine-tuned on specific medical imaging tasks with relatively modest data requirements. The authors believe that the availability of a large-scale, three-dimension–capable, and extensively pretrained transformer model would be highly beneficial to the medical imaging and research community. In this article, authors discuss the challenges and obstacles of training a very large medical imaging transformer, including data needs, biases, training tasks, network architecture, privacy concerns, and computational requirements. The obstacles are substantial but not insurmountable for resourceful collaborative teams that may include academia and information technology industry partners.

© RSNA, 2022

Deep learning methods dominate large areas of medical imaging analysis. At the forefront are convolutional neural networks (CNNs) such as ResNets (1) or DenseNets (2) for classification tasks or the famous U-Net for segmentation tasks (3). While progress is still being made, one has to attest that—for example, for segmentation, since the introduction of the three-dimensional (3D) U-Net—while there have been innumerable adaptations to specific tasks, the generic technological core has had only modest innovations, if any at all. This can be exemplarized by the no-new-U-Net (nnU-Net) (4), which has won competitions in multiple tasks years later with a basically "plain vanilla" U-Net by focusing on components outside the model architecture, like data preprocessing and model ensembling. While the segmentation performance of CNNs on average appears to be sufficient for many use cases, there is still a large gap compared with human performance, as can be seen in the sometimes obvious mistakes these models make, especially if there is a domain shift. For example, complete organs are ignored when presenting noncontrast CT data to networks trained on contrast CT data, while humans would have minimal difficulty in adapting to this change in appearance (5). Therefore, there is certainly a need for improvement.

One hope was to overcome these limitations with larger amounts of training data. But deep learning for medical imaging faces the specific challenge that, depending on the task, it may be very expensive or difficult to obtain well-curated labeled data. At the same time, the expectations for accuracy in real-world scenarios are exceedingly high. A way to improve the situation is to use transfer learning with networks that are pretrained on (nonmedical) common image collections, such as Imagenet, or video data (6), to have a component of 3D (or more specifically, two dimensions [2D] + time). This does have benefits for performance, but intuitively, classifying common objects such as "cat," "dog," "chair," and so forth does not convey the intricacies of human 3D anatomy as captured by medical imaging devices. Attempts to create CNNs that are pretrained on larger amounts of medical imaging data have been made (7,8) and show some success. While pretraining on 3D medical data showed the optimal performance, it should be noted that the benefit of pretraining on medical data instead of generic 2D images appears to be limited (8). This likely points to an inherent limit of CNN architectures. Ideally, a neural network architecture would benefit from transfer learning on several levels, from low-level features such as generic edges, to more specific features such as certain textures, to higher-level features like shapes of human organs. The finding that pretraining with generic images yields similar performance compared with pretraining with medical images suggests—although we acknowledge that there is a paucity of evidence—that pretraining mainly pertains to low-level nonspecific features or is at least unable to capitalize on mid- or higher-level morphologic features.

Transformer Networks

In the meantime, while the medical imaging research community employed CNNs and variations, in the natural language processing (NLP) field, a major shift has taken place with the introduction of transformer networks (9). Transformers have not only increased the performance of NLP tasks across the board but have also

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

Abbreviations

CNN = convolutional neural network, GPU = graphics processing unit, NLP = natural language processing, nnU-Net = no-new-U-Net, 3D = three-dimensional, 2D = two-dimensional

Summary

Transformer networks may replace convolutional neural networks but need very large amounts of training data. Training of a large-scale three-dimensional medical imaging transformer model could possibly reduce the amount of data needed to train and fine-tune medical imaging models.

Key Points

- Training of a large-scale three-dimensional medical imaging transformer model could possibly reduce the amount of data needed to train and fine-tune medical imaging models.
- Challenges and obstacles of training a very large universal medical imaging transformer network include data needs, biases, training tasks, network architecture, privacy concerns, and computational requirements.
- These obstacles are substantial but not insurmountable for resourceful collaborative teams that may include academia and information technology industry partners.

Keywords

Computer-aided Diagnosis (CAD), Informatics, Transfer Learning, Convolutional Neural Network (CNN)

enabled tasks that were previously considered impossible with traditional methods, such as, for example, learning language translation from scratch. Transformer networks have inner workings that are fundamentally different from CNNs and are based on sequence processing (10).

A major difference compared with CNNs is that transformer networks have a strong inherent attention mechanism. While in a CNN a certain neuron only takes inputs from adjacent neurons in the previous layer, in a transformer the inputs can be taken from anywhere in the prior layer. Which features are used for processing is determined by a learned attention mechanism. Depending on the input data each layer can select certain parts of the input for further processing. How is this possible? At the core, a rather complicated neural network mechanism mimics the key, query, value idea from traditional databases, where a value is retrieved from a table based on a query matching to a key. This attention concept is shown in Figure 1. Figure 1A shows attention in a text model where the transformer correctly identifies the link between the words *transformers* and *they* in the attention maps. Figure 1B (from Cordonnier et al [11], with permission) shows an example of a visual transformer where attention can be focused on adjacent pixels (Fig 1B, upper two examples) or focus on a more distant portion of the inputs (Fig 1B, lower example). This shows that a transformer can easily model local and distant (long-range) relationships in the presented data, which is challenging for CNNs, as they always have a highly local context. A complete description of the transformer architecture is outside of the scope of this article, and we refer the reader to Vaswani et al (12) for further information.

Empirically, transformers require very large amounts of training data to reach full potential. As labeled data are scarce, it is preferable to use a self-supervised task for pretraining and then fine-tune on a specific task with available labels. In NLP, an example for a self-supervised task would be to predict a masked word in a sentence. It is obviously very simple to mask a word in a sentence, but it can be very difficult to guess this word. With this task, transformers can be trained on very large-scale unlabeled datasets, such as the complete text of Wikipedia plus all full texts available in PubMed. The term *foundation model* has been recently coined to describe large, extensively pretrained networks (9). The pretrained network can then be fine-tuned on a specific task, with a low amount of annotated data.

It did not take long until this concept was transferred to imaging tasks where an image is divided into small patches and the patches are treated as "words"; therefore, an image correlates with a sequence of words, as was shown in a seminal publication from Google Brain, *An Image is Worth 16 × 16 Words* (13). Vision transformers now outperform CNNs on a multitude of tasks, but they still require exceedingly large amounts of data to reach their true potential (10). Vision transformers have fascinating properties that overcome major shortcomings of CNN architectures, namely the inability to model long-distance spatial relationships and the overreliance of CNNs on textures with a weakness in modeling shapes (14). Vision transformers, by design, can easily model long-distance visual relationships, which is shown in Raghu et al (10). In addition, the ability to function more on the basis of shapes rather than on textures is also highly attractive and promises better generalizability and robustness (15). It should be emphasized that the ability to resolve these shortcomings of CNNs was elusive for many years or required task-specific tweaks. Examples of transformer-based and nnU-Net segmentations are shown in Figure 1C (based on model from Tang et al [16]).

The described models and training datasets require large computational resources; for example, the equivalent of 2500 days of computation on an advanced neural network accelerator (TPUv3) was used for the Google ViT H14 model (greatly accelerated by parallel training). After training, the complex models can potentially be simplified (distilled [17]) and then can be fine-tuned on a new imaging task with much lower resource requirements, down to a single graphics processing unit (GPU), which makes them accessible to virtually all machine learning researchers. Hybrid models combining transformers and CNNs have been developed and show promising results in medical imaging applications (18–20) (Fig 2). The advantages of this hybrid approach compared with a pure transformer approach are higher computational efficiency and the ability to include skip connections, which have been shown to be highly beneficial in CNN architectures. The transformer encoder block enables the model to have the full potential of the transformer attention mechanism to model long-range dependencies, which is not present in pure CNNs.

One could argue that the next logical step certainly would be the training of a large-scale 3D medical imaging transformer model, which then could increase performance and possibly reduce the amount of data needed to train or fine-tune medical imaging models across new tasks.

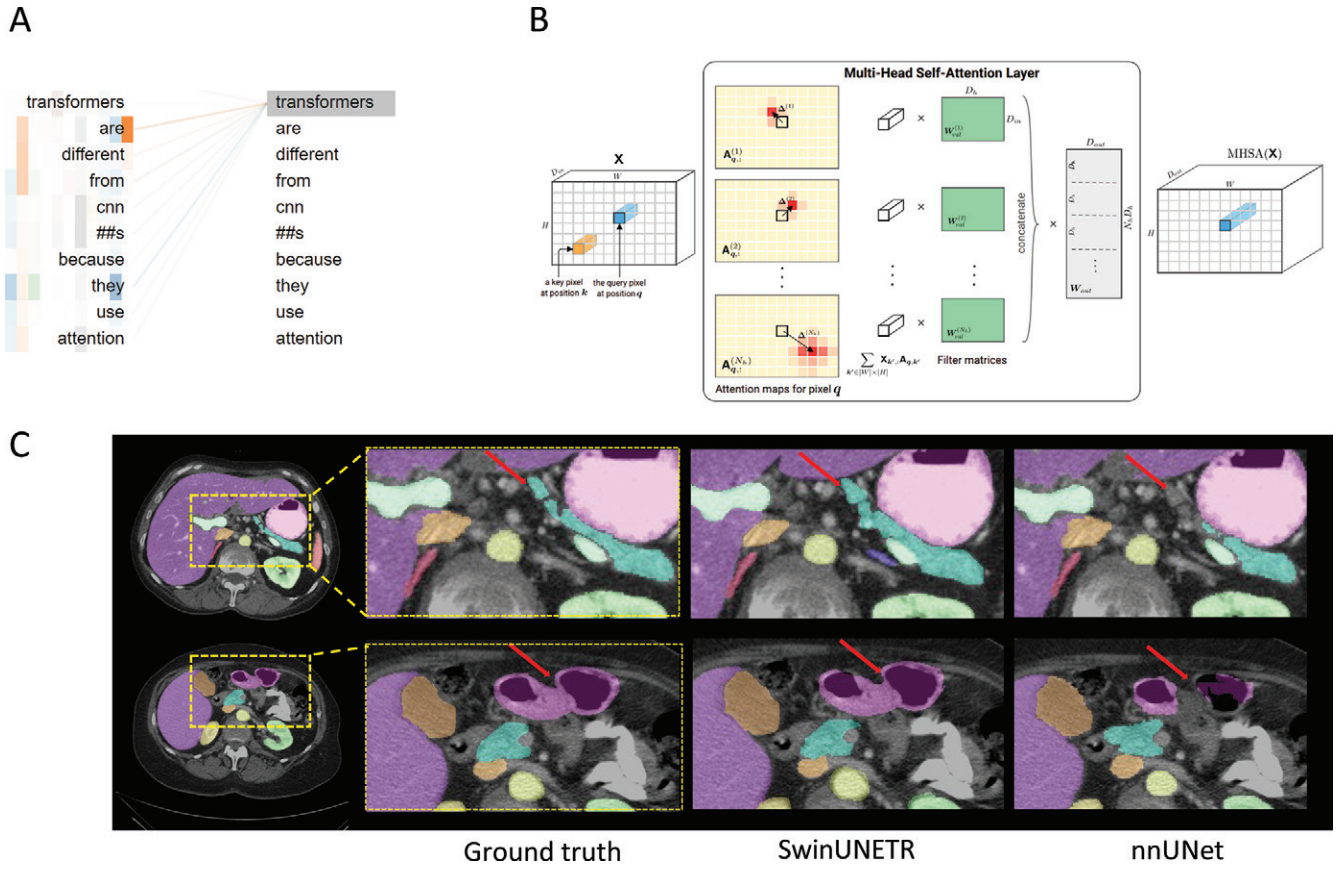


Figure 1: Visualization of attention maps in **(A)** a natural language processing transformer and **(B)** a visual transformer. The attention map in the text transformer shows the relationship between the word *transformers* and the adjacent verb *are*, as well as the pronoun *they* later in the sentence. A similar concept is shown in **B** in a visual transformer, where attention maps can relate to nearby pixels (upper two examples) or can be remote and shifted (bottom example, reproduced with permission from reference 11). **(C)** Example transformer-based and no-new-U-Net (nnU-Net) segmentations of abdominal CT images show the highest performance in a pretrained transformer (SwinUNETR) for structures that are traditionally difficult to segment, for example, pancreas or stomach (red arrows are pointing to specific areas for which SwinUNETR shows superior results) (SwinUNETR described in reference 16; image credit, Yucheng Tang, personal communication). CNN = convolutional neural network, MHSA = multihead self-attention, UNETR = U-Net transformer.

What Are the Challenges and Obstacles in This Task?

Data

A large amount of data are required, and the entirety of openly available datasets may not be sufficient or optimal for the training. This issue may need data curation, labeling, and sharing of currently siloed very large data collections. A method to overcome this issue may be federated learning, which could be a way to avoid the need for centralized data in the future (21,22). While data scarcity is not an issue specific to transformer networks, it has been shown that transformers require larger amounts of data compared with CNNs to reach optimal performance, and therefore, data scarcity becomes more important with large transformer models.

Biases

The training data very likely in some form contain societal and socioeconomic biases, starting with over- or underrepresentation of certain groups within the data, different imaging techniques, and others. Language models have been shown to contain extreme biases (23). It should be strongly emphasized that each resulting fine-tuned network has to be carefully analyzed for biases in the specific application task, and measures should be taken to counteract any biases, for example, by ensuring evenly sampled data and performance across all groups. Nevertheless, this is an unresolved issue in medical AI in general, and ongoing efforts are needed to hopefully improve this situation.

Training Task

Simple training tasks that do not require human-generated labels have to be devised. For example, Image GPT (24) was trained without labels in a fashion similar to the NLP masked word task, where instead of a word in a sentence, a part of an image was masked, to be predicted by the network. Additionally, medical data have some specific properties that could be used to generate automatic labels. For example, at MRI, multiple sequences are available that show different properties of the imaged tissues, and dual-energy CT provides iodine maps. Detailed suggestions for automatically generated tasks for self-supervised learning in medical images were described by Zhou and colleagues (8). Often, radiology reports are available, and simplified (and probably imperfect) labels can be automatically generated from the reports. While some of these methods

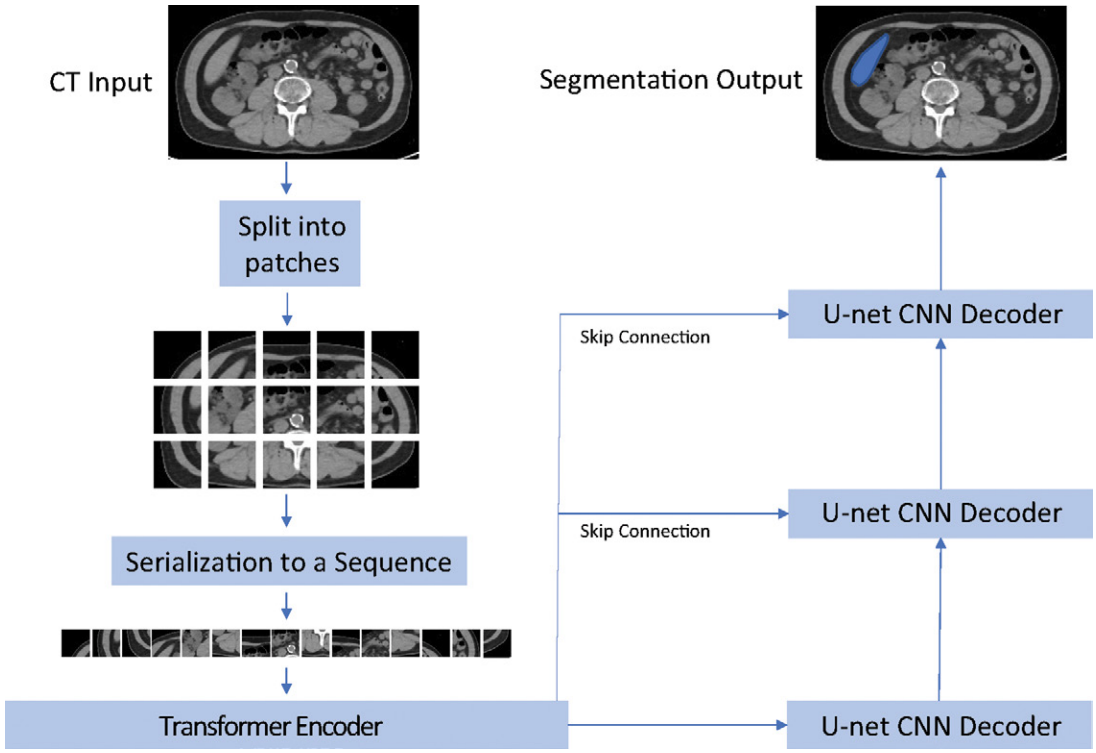


Figure 2: Example of a U-Net–transformer hybrid (UNETR, figure based on ideas from reference 18). The input image volume is split into a large number of small patches that are then serialized to a one-dimensional sequence (with the addition of a localization token to preserve spatial relation information, not shown). A transformer model is applied to this sequence, analogous to natural language processing models, and functions as an encoder. The encoded compressed representation is then decoded via multiple CNN decoder modules with skip connections, similar to the decoding steps of a traditional U-Net, and a loss is applied for the resulting segmentation. The model can be trained end to end. CNN = convolutional neural network.

may not apply to medical image segmentation tasks, nonexpert crowdsourcing has the potential to make human segmentations available on a large scale (25). In summary, a multitude of tasks can be envisioned and then tested for efficacy in the training of a vision transformer.

Network Architecture

The current vision transformers are designed for small 2D images. Medical imaging networks would ideally process complete 3D CT or MRI volumes, but smaller 3D sizes are certainly an acceptable interim step. Hybrid models involving CNN and transformer components already enable the processing of usable 3D resolution with current hardware. It should be emphasized that at this time it is unclear if the level of generalization and performance improvement that is seen in NLP applications with large networks and large datasets will also emerge in medical imaging applications, but looking at the results on unsupervised learning with imaging transformers, this appears to be likely (26). Also, it is unclear if the more efficient hybrid CNN-transformer networks will benefit from very large dataset pretraining in the same way. Many questions regarding the optimal architecture are unresolved, and we are truly just at the beginning of this exploration.

Privacy Concerns

Depending on the data used, there may be privacy concerns, as a very large model may "remember" certain specific images, and it cannot be excluded that such data can be extracted. Of note, this is also an issue with CNNs, although it is certainly exacerbated in transformers because of the larger amounts of data used. Ideally, the training data should therefore be fully anonymized and/or already openly available. One should be aware that data from specific body parts are difficult to anonymize. For example, facial characteristics may be visible on MR images of the brain (27), which can lead to reidentification. Technologies for facial anonymization are available or in development, including the removal of facial features or the use of generative adversarial networks (28).

Compute

It is likely that full-scale training will require very large amounts of computation involving tens to thousands of GPUs, but it should be emphasized that this is only needed during development. Once the network is trained, it can be hopefully fine-tuned for other tasks with easily available resources. Technologies are in development that may very substantially reduce the required computation: for example, the use of teacher-student models (26). In addition, some of the issues brought up can be analyzed, worked on, and improved using exploratory smaller models without the need for large-scale computation.

The obstacles for training a very large universal medical imaging transformer network discussed above are important but are not insurmountable for resourceful collaborative teams, which may include academia and information technology industry partners. If successful, the promise is that computer vision for

medical imaging could make, once again, a quantum leap, similar to what we already see in the field of NLP. Exciting times in the field are upon us. We have no doubt that these new developments will in the longer term positively affect patient care.

Author contributions: Guarantor of integrity of entire study, **V.S.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, all authors; and manuscript editing, all authors

Disclosures of conflicts of interest: **M.J.W.** Grant or contract from the American Heart Association (no. 18POST34030192); consulting fees from Segmed; support from Segmed for attending meetings and/or travel; stock or stock options in Segmed. **H.R.R.** Employed by NVIDIA. **V.S.** No relevant relationships.

References

1. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 27–30, 2016. Piscataway, NJ: IEEE, 2016; 770–778.
2. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017. Piscataway, NJ: IEEE, 2017; 2261–2269.
3. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science, vol 9901. Cham, Switzerland: Springer, 2016; 424–432.
4. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–211.
5. Sandfort V, Yan K, Pickhardt PJ, Summers RM. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Sci Rep 2019;9(1):16884.
6. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, December 7–13, 2015. Piscataway, NJ: IEEE, 2015; 4489–4497.
7. Chen S, Ma K, Zheng Y. Med3D: Transfer Learning for 3D Medical Image Analysis. arXiv 1904.00625 [preprint] https://arxiv.org/abs/1904.00625. Posted April 1, 2019. Accessed August 24, 2021.
8. Zhou Z, Sodha V, Rahman Siddiquee MM, et al. Models genesis: generic autodidactic models for 3D medical image analysis. In: Shen D, Liu T, Peters TM, et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11767. Cham, Switzerland: Springer, 2019; 384–393.
9. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv 2108.07258 [preprint] https://arxiv.org/abs/2108.07258. Posted August 16, 2021. Accessed September 7, 2021.
10. Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. Advances in Neural Information Processing Systems 34 (NeurIPS 2021).
11. Cordonnier JB, Loukas A, Jaggi M. On the relationship between self-attention and convolutional layers. arXiv 1911.03584 [preprint] https://arxiv.org/abs/1911.03584. Posted November 8, 2019. Accessed January 15, 2022.
12. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, et al, eds. Advances in Neural Information Processing Systems 30 (NeurIPS 2017).
13. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations. 2021.
14. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations. 2019.
15. Naseer M, Ranasinghe K, Khan S, Hayat M, Khan FS, Yang MH. Intriguing properties of vision transformers. arXiv 2105.10497 [preprint] https://arxiv.org/abs/2105.10497. Posted May 21, 2021. Accessed August 25, 2021.
16. Tang Y, Yang D, Li W, et al. Self-supervised pre-training of swin transformers for 3D medical image analysis. arXiv 2111.14791 [preprint] https://arxiv.org/abs/2111.14791. Posted November 29, 2021. Accessed January 15, 2022.
17. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning. 2021; PMLR 139:10347–10357.
18. Hatamizadeh A, Yang D, Roth H, Xu D. UNETR: transformers for 3D medical image segmentation. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, January 3–8, 2022. Piscataway, NJ: IEEE, 2022; 1748–1758.
19. Petit O, Thome N, Rambour C, Soler L. U-Net transformer: self and cross attention for medical image segmentation. arXiv 2103.06104 [preprint] https://arxiv.org/abs/2103.06104. Posted March 10, 2021. Accessed August 24, 2021.
20. Chen J, Lu Y, Yu Q, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv 2102.04306 [preprint] https://arxiv.org/abs/2102.04306. Posted February 8, 2021. Accessed September 2, 2021.
21. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. Radiology 2020;295(1):4–15.
22. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3(1):119.
23. Next chapter in artificial writing. Nat Mach Intell 2020;2(8):419.
24. openai. GitHub - openai/image-gpt. https://github.com/openai/image-gpt. Accessed August 24, 2021.
25. Bhatter P, Frisch E, Duhaime E, Jain A, Fischetti C. Diabetic retinopathy detection using collective intelligence. J Sci Innov Med 2020;3(1):1.
26. Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, October 10–17, 2021. Piscataway, NJ: IEEE, 2021; 9650–9660.
27. Schwarz CG, Kremers WK, Therneau TM, et al. Identification of anonymous MRI research participants with face-recognition software. N Engl J Med 2019;381(17):1684–1686.
28. Parker W, Jaremko JL, Cicero M, et al. Canadian Association of Radiologists White Paper on De-Identification of Medical Imaging: Part 1, General Principles. Can Assoc Radiol J 2021;72(1):13–24.