Medical Transformers

Ramiz Dundar
ramiz.dundar@boun.edu.tr

Ahmet Emir Kocaaga
ahmet.emir.kocaaga@boun.edu.tr

Abstract

Transformers [29] are a widely adopted technique in the natural language processing research community due to their powerful attention and parallelization capabilities. However in the computer vision scene, before vision transformer (ViT) [15], pure applications of transformers on images often are underperformed compared to convolutional neural networks or hybrid models (CNNs and transformers together). Vision transformer (ViT), although a relatively new approach, has the state of the art results in image classification. In this paper, 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) [28]. Further analyses and comparisons between different applications of vision transformers are given.

1. Introduction

Transformer encoder-decoder models [29] have been the state-of-the-art approach in many natural language processing (NLP) applications due to their computational efficiency and scalability. They use a fully attention-based approach shown to outperform traditional sequence models based on recurrent architectures [17] in several applications, and have become the de-facto standard for several NLP tasks. The Transformer architecture benefits from a deep stack of self-attention layers [8], layer normalization [1] and positional encodings [29, 16]. Even though originally proposed as an encoder-decoder model for machine translation task [29], both encoder and decoder components have been successfully employed to large-scale models such as BERT [14] and GPT [23] as a robust architecture for learning self-supervised representations of text.

This success in NLP applications have been inspired the adaptation of self-attention based architectures in computer vision [24, 30]. However, a straightforward application of self-attention to images does not scale to realistic applications due to its quadratic cost in number of pixels. Since the memory requirement in Transformers linearly increases in terms of number of tokens in the sequence, most of the Transformer-based models for computer vision applications use various approximations such as applying self-attention to local neighborhoods [21], using sparse factorizations [9] or combining self-attention with convolutional neural networks [2].

Recently, Vision Transformer (ViT) [15] is proposed as the first study that uses Transformers with global self-attention on full-sized images and gained significant attention due to its state-of-the-art results. Their model outperforms convolutional neural networks in image classification tasks while requiring significantly less computational resources to train. Their key insight is splitting an image into patches, and providing a sequence of linear embeddings of these patches as an input to a Transformer. In other words, image patches are treated as tokens the same way as words in NLP applications.

Inspired by the success of the Vision Transformer on image classification task, we explore application of Transformers on the X-ray classification for certain diseases. Similar to Vision Transformer, we follow the strategy of splitting an image into patches and provide a sequence of their corresponding linear embeddings as an input to Transformer encoder.

For medical images we use Chexpert [18] dataset. It has more than two hundred thousand images. Each study in dataset has two images taken from side and front and labeled with the 14 clinically important and prevalent observations. We try to predict between 5 diseases selected based of clinical importance among these 14 observations

We also propose a strided vision transformer approach which modifies this batching strategy. Our strided approach aims to better capture inherent relation between different patches of the image. It does this by using patches which are overlapping with each other. Doing this we expect to have better attention mechanism than what is proposed previously. Experimental results on Chexpert[18] also align with what we expect and strided approach gain considerable improvements on top of the existing ViT model.

One recent study also shown that using distillation reduced the effort to train [28] for Vision Transformers [15]. Following that we explored usage of distillation on Vision

Transformers on medical domain.

To the best of our knowledge, this is the first study that tackles a transformer only based approach with a global self-attention on full-sized images in X-rays.

The rest of this paper is organized as follows. In Section 2, we discuss related work on transformers and X-ray classification. In Section 3, we describe our model. In Section 4, we compare our method with several baselines and demonstrate the effectiveness of our model both via quantitative and qualitative experiments. Section 5 concludes the paper.

2. Related Work

In this section, we review related work in transformers and image-to-image translation task.

2.1. Transformers

Transformers are proposed for machine translation task [29] and have been extensively used in a variety of NLP tasks as well as serving as a basis for popular large-scale models such as BERT [14] and GPT [22]. BERT is prominent in the sense that it can be applied to a variety of tasks, such as predicting next sentence of a given sentence while each task does not require significant change of the architecture. The performance of GPT stems from its task-specific discriminative fine-tuning which lets GPT to be applicable to many natural language understanding tasks without changing the architecture substantially. GPT has been improved in recent years and the most recent model, GPT-3 [3], scales up the previous models while using an autoregressive language model.

One limitation of transformers is that they are often restricted to fixed-length context. In [13], Transformer-XL has been proposed for addressing this issue. It not only allows variable-length but also maintains temporal coherence. Later, Transformer-XL inspired the XLNet [33] architecture. XLNet provides a generalized autoregressive pretraining method and resolves pretrain-finetune discrepancy with BERT. Another limitation of transformers is their cost. Although large transformers have been shown to obtain state-of-the-art results, training complexity if high on, especially, long sequences. It has been demonstrated in [20] that transformers can be used on long sequences with small memory usage while performing as good as the less memory efficient transformers.

Recently there has been growing interest towards adapting transformers for computer vision tasks. One line of research aims to combine self-attention with convolutional neural networks (CNNs) for a variety of tasks including semantic segmentation [34], object detection [4], and text-to-video generation [27]. [11] shows the expressiveness of self-attention layers by comparing them with convolutional layers. They demonstrate that a multi-head self-attention layer can express any convolutional layer when the former has sufficiently many heads.

Another line of research aims to directly apply Transformers for computer vision tasks, however a straightforward application of self-attention to images does not scale to realistic applications due to its quadratic cost in number of pixels. Therefore, various approximations were proposed in the past such as applying self-attention to local neighborhoods [21], or using sparse factorizations to approximate to global self-attention [9]. [10] focus on reducing the time and space complexity of attention by computing soft attention on small chunks of input. They obtain the chunks by adaptively dividing the input. [31] uses a block-local attention that has high training speed on TPUs. This allows them to use three-dimensional data (eg. video) efficiently. [30] decreases the time complexity by representing 2D self-attention as 1D self-attention. Hence, this approach opens a road to apply attention on a larger region.

The first direct application of Transformers with global self-attention to full-sized images is proposed in [15] which splits an image into patches and feeds the sequence of linear embeddings as an input to the Transformer Encoder. They demonstrated that using a simple-patch based approach outperforms state-of-the-art convolutional networks. Their exploitation of ample data resources leads to state-of-the-art results.

Image GPT [6] trains a transformer that predicts pixels while not including structure of the image. Their model is trained on low-resolution images and learns image representations in an unsupervised manner. GAN-BERT [26] trains BERT for generating PET images from MRI images and their methods is easily applicable as it can scale. They chose the Next Sentence Prediction (NSP) module as their GAN discriminator.

We observed that the prior research on transformers haven't touched upon X-ray classification yet.

2.2. Transformers in the medical area

In the medical domain, there is prior work done with transformers. This work mostly focuses on medical report classification or report generation (NLG). One such example is Generating Radiology reports via Memory-driven transformer [7], which aimes to generate radiology reports automatically from a given chest x-ray. In this paper, images are fed into convolutional neural networks to extract features. Extracted features are used by transformers to generate radiology reports. Furthermore, transformers are enhanced with a unit called Relational Memory which acts as a memory between images for transformer architecture. The experiments are conducted on two radiology report datasets, MIMIC-CXR and IU X-Ray.

Another example where a automated medical report generation framework is provided is Reinforced Transformer

[32]. In this paper, a hierarchical Transformer-based medical report generation method is proposed. In this method, an encoder-decoder mechanism consisting of transformers is used and thereby medical reports are generated. The experiments of this paper is also conducted on IU X-Ray radiology report dataset.

A different application of transformers in medical domain is CGMVQA [25], which is a medical visual question answering framework that aims to assist doctors in clinical analysis. In this paper, features extracted by ResNet152 are used by transformers.

Recently, after the success of vision transformer (ViT) [15] on image classification tasks, hybrid convolutional neural network - transformer architectures are created for medical image classifications tasks. One such example is TransMed [12]. In this paper, a convolutional neural network is used as a feature extractor and the extracted features from a medical image is given to a transformer to classify medical images.

Another interesting application of vision transformer (ViT) [15] on medical domain is Vit-V-Net [5]. Vit-V-Net is also a convolutional neural network - transformer architecture, which is proposed for medical image registration. Convolutional newural networks are used as a encoder before vision transformer (ViT) [15] layer in Vit-V-Net.

Lastly, pure Transformer architectures without using any convolutional layers are proposed for medical image segmentation. In the paper Convolution-Free Medical Image Segmentation using Transformers [19] a pure transformer framework is used for image segmentation. The network splits 3D Images into smaller 3D patches, computes 1D embeddings for each of these patches and these embeddings are given to a self-attention layer.

3. Method

We propose a Transformer-based image to image translation model. We first review the recently proposed vision transformer (ViT) model [15] which uses a patch-based Transformer encoder. Then, we describe how we extend their framework by introducing a strided batch geneation approach. Lastly we do the same for distilled vision transformer (DViT) model [28] which uses distillation tokens.

3.1. Dataset

For this project, we use the Chexpert [18] dataset. Chexpert is a large chest X-ray image dataset with 224,316 chest radiographs of 65,240 patients. Each X-ray image is either taken from the side (lateral) or taken from the front (frontal). Also, the frontal image category is divided into Posterior-Anterior (PA) or Anterior-Posterior (AP).

[18] is a dataset for image classification. There are no radiology reports written but only image labels. Besides age and gender information, for each radiology report, 14 observations based on the clinical importance selected, and then using a rule-based automated extractor, labels were assigned to each report. We trained our model with these labels.

Automated label extraction outputs one of the following for each report: 1 for the positive, 0 for the negative, 0 for the uncertain, and lastly blank for the unmentioned.

3.2. Vision Transformers (ViT)

The original transformer framework receives input as a 1D sequence. In order to handle 2D sequences, [15] proposed a patch-based that splits an image into patches and feeds the sequence of linear embeddings as an input to the Transformer Encoder (see Figure **??**). In other words, 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 $\mathbf{x_p} \in \mathbb{R}^{N \times P^2 \dot{C}}$ 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. Position embeddings in Vision Transformers follows standard 1D position embeddings, and added to the patch embeddings in order to retain positional information.

Vision Transformers follows the original design of Transformer encoder [29] closely, and uses multi-headed self attention and feed forward blocks as follows:

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

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

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

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

where $\mathbf{MSA}$ 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.

3.3. Strided ViT (SViT)

When transformers are used in NLP tasks, the attention mechanism produces results for combinations of each word in the text. Such an attention mechanism allows the transformer to extract all the relations between those words, therefore they produce highly accurate results.

While the building blocks of texts are words, those are pixels for images. So, the first thing that comes to mind is creating an attention mechanism for combinations of each pixel, but this is not possible with a usual GPU since this mechanism needs a huge amount of memory. Thus, Vision Transformer (ViT) [15] proposes a patch-based attention mechanism. But this method loses some features that

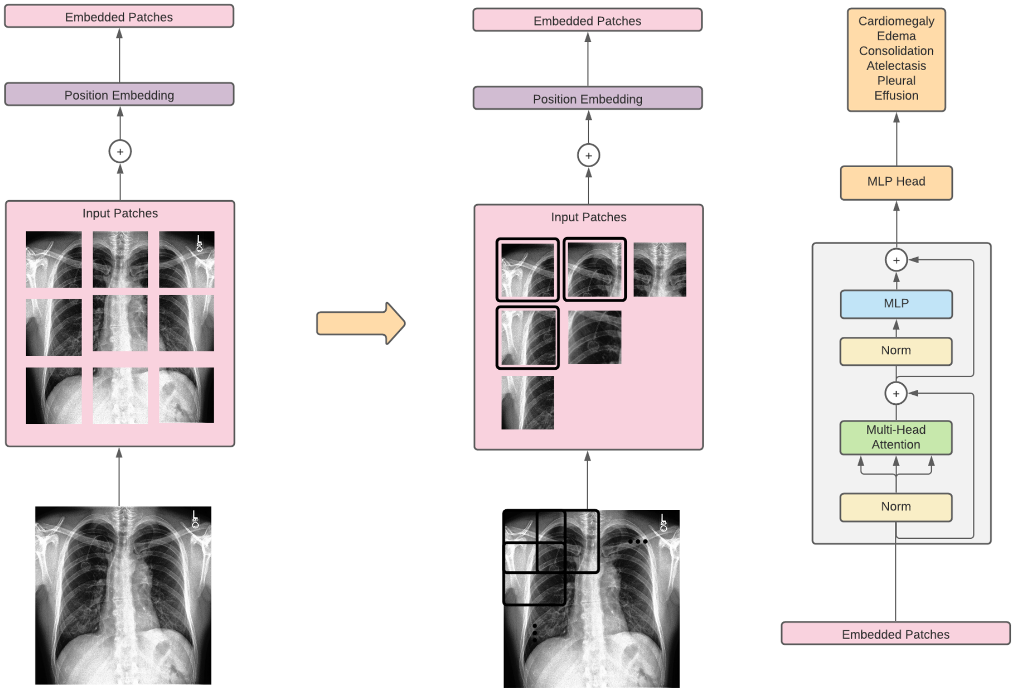

Figure 1. Strided Vision Transformer

can be extracted from the image since the maximum attention capacity is not achieved.

We propose a "strided" approach both to eliminate this problem and to provide memory efficiency (see Figure 1). In this approach, we again split the image into the patches, but this time those parts of those patches overlap. For example, assume that the right border of the leftmost patch is on the Nth pixel. The next patch, which is on the right side of this patch, starts from the N+1th pixel in Vision Transformer (ViT) [15]. But in our approach, it starts from N-x th pixel, so it produces more attention by using NxN patches.

3.4. Distilled ViT (DViT)

In a recent paper, [28] showed that using distillation tokens to distill information from an already trained convolutional neural network, reduced the time and resources needed to train the vision transformer.

At its core distill ViT takes additional two hyperparameters. These two are used for adjusting how much of the loss will come through distillation. In this environment, ViT acts as a student learns from the convolutional neural network, which we will refer to as the teacher model.

$$teacherLoss = teacherLoss * temperature \tag{5}$$

$$loss = studentLoss * \alpha + teacherLoss * (1 - \alpha) \tag{6}$$

We tested how distillation performed in the medical domain. And we also applied a strided patching approach to vision transformers using distillation.

4. Results and discussion

In our experiments, we tried four different approaches to increase our medical transformers' prediction accuracy. For all of our experiments, we did the implementation using PyTorch in Python.

4.1. ViT

Our first approach was directly using Vision Transformer (ViT) [15] in the Chexpert dataset [18]. The result we got was a mean AUC score of 0.858 and it is below the Stanford Baseline for the Chexpert dataset. Since Vision Transformer produces better results with larger datasets, this was expected. On the other hand, this score was promising and it showed that Vision Transformer (ViT) [15] is applicable to medical domain and open to further improvements with modifications.

4.2. Strided ViT

Since Vision Transformer (ViT) [15] follows a path-based approach in the attention layer, some information between these patches is lost. To prevent this, we followed a strided approach.

	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

Table 1. Mean AUC Results

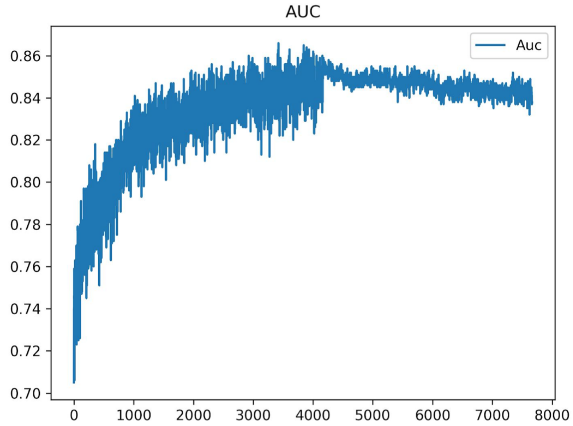

Figure 2. ViT AUC Score

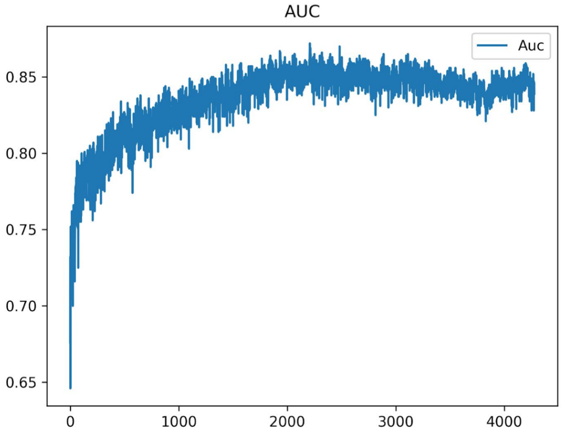

Figure 3. DS-ViT AUC Score

The result we got with Strided VIT is better than Vision Transformer (ViT) [15]. The mean AUC score is 0.867. So, it showed a progress more than 1% in terms of accuracy and proved that it has potential. By this result, we can conclude that the extra attention we use in strided approach is beneficial in terms of model accuracy.

4.3. Distillable ViT

As stated above, Vision Transformer (ViT) [15] needs a huge amount of input to produce better results. But Distillable Vit [28] can learn more with less data. So, since our dataset is comparatively small, we tried also Distillable Vit [28]. For distillation, a pre-trained model is needed and we used a pre-trained DenseNet121, which was also trained on the Chexpert dataset [18].

Distillable Vit [28] was able to produce better results. Using the distillation tokens, Vision Transformer (ViT) [15] was able to learn better and as a result, we got a mean AUC score of 0.87.

4.4. Distillable, Strided ViT

In this approach, we combined distillation and strided approach. As expected, it increased the accuracy of the model and got a mean AUC score of 0.88. As a result, by combining distillation and strided approach, we have managed to increase accuracy of Vision Transformer (ViT) [15] by more than 2.5%.

5. Conclusions

In this work, we developed an image classification model for medical X-ray images using the Chexpert dataset [18]. We also used vision transformers for the first time in the medical domain. Our proposed model used the combined architecture of convolutional neural networks and the vision transformer.

We developed strided vision transformers where images are given into vision transformers in a strided fashion. In our experiments, we observed that we gained more than one percent increase in mean AUC score using the strided version.

Chext X-rays in the Chexpert dataset [18] are fundamentally different from the images that the original Vision Transformer [15] trained on and shows a state of the art prediction accuracy. Furthermore, Chexpert dataset [18] is a relatively small dataset and the Vision Transformer [15] needs a large dataset in order to show its strength. Despite these facts, Distillable Vit [28] and our strided aprroach showed a promising performance on the Chexpert dataset [18].

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3286–3295, 2019.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

[5] Junyu Chen, Yufan He, Eric C. Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration, 2021.

[6] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 1, 2020.

[7] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer, 2020.

[8] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

[9] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[10] Chung-Cheng Chiu and Colin Raffel. Monotonic chunkwise attention. *arXiv preprint arXiv:1712.05382*, 2017.

[11] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2019.

[12] Yin Dai and Yifan Gao. Transmed: Transformers advance multi-modal medical image classification, 2021.

[13] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

[19] Davood Karimi, Serge Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using transformers, 2021.

[20] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[21] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.

[22] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[24] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[25] Fuji Ren and Yangyang Zhou. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, PP:1–1, 03 2020.

[26] Hoo-Chang Shin, Alvin Ihsani, Swetha Mandava, Sharath Turuvekere Sreenivas, Christopher Forster, Jiook Cha, and Alzheimer's Disease Neuroimaging Initiative. Ganbert: Generative adversarial networks with bidirectional encoder representations from transformers for mri to pet synthesis. *arXiv preprint arXiv:2008.04393*, 2020.

[27] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019.

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[30] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. *arXiv preprint arXiv:2003.07853*, 2020.

[31] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019.

[32] Yuxuan Xiong, Bo Du, and Pingkun Yan. *Reinforced Transformer for Medical Image Captioning*, pages 673–680. 10 2019.

[33] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

[34] Chengju Zhou, Meiqing Wu, and Siew-Kei Lam. Ssa-cnn: Semantic self-attention cnn for pedestrian detection. *arXiv preprint arXiv:1902.09080*, 2019.