CmpE 492: Medical Transformers

Ramiz Dündar

Ahmet Emir Kocaağa

1. Introduction and Motivation

Transformer encoder-decoder models Vaswani et al. (2017)[1] 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 Hochreiter and Schmidhuber(1997)[2] 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 Cheng et al.(2016)[3], layer normalization Ba et al. (2016)[4], and positional encodings Vaswani et al. (2017)[1]; Gehring et al. (2017)[5]. Even though originally proposed as an encoder-decoder model for machine translation task Vaswani et al. (2017)[1], both encoder and decoder components have been successfully employed in large-scale models such as BERT Devlin et al. (2018)[6] and GPT Radford et al. (2019)[7] as a robust architecture for learning self-supervised representations of text. This success in NLP applications has inspired the adaptation of self-attention-based architectures in computer vision Ramachandran et al. (2019)[8]; Wang et al. (2020)[9]. However, a straightforward application of self-attention to images does not scale to realistic applications due to its quadratic cost in the number of pixels. Since the memory requirement in Transformers linearly increases in terms of the 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 Parmar et al. (2018)[10], using sparse factorizationsChild et al. (2019)[11] or combining self-attention with convolutional neural networks Bello et al. (2019)[12]. Recently, Vision Transformer (ViT) Dosovitskiy et al. (2020)[13] 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 the 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 the Transformer encoder. To the best of our knowledge, this is the first study that tackles a transformer-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 Transformer 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. State of the Art

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

2.1 Transformers

Transformers are proposed for machine translation tasks Vaswani et al.(2017)[1] 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 Devlin et al. (2018)[6] and GPT Radford et al.(2018)[14]. BERT is prominent in the sense that it can be applied to a variety of tasks, such as predicting the next sentence of a given sentence while each task does not require a significant change of the architecture. The performance of GPT stems from its task-specific discriminative fine-tuning which lets GPT apply 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 Brown et al. (2020)[15], scales up the previous models while using an autoregressive language model.

One limitation of transformers is that they are often restricted to fixed-length contexts. In Dai et al. (2019)[16], 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 Yang et al. (2019)[17] architecture. XLNet provides a generalized autoregressive pretraining method and resolves pre-train-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 is high on, especially, long sequences. It has been demonstrated in Kitaev et al. (2020)[18] that transformers can be used on long sequences with small memory usage while performing as well as the less memory efficient transformers.

Recently there has been growing interest in 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 Zhou et al. (2019)[19], object detection Carion et al. (2020)[20], and text-to-video generation Sun et al. (2019)[21]. Cordonnier et al. (2019)[22] 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 the number of pixels. Therefore, various approximations were proposed in the past such as applying self-attention to local neighborhoods Parmar et al. (2018)[10]or using sparse factorizations to approximate global self-attention Child et al. (2019)[11]. Chiu and Raffel (2017)[3] 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. Weissenborn et al. (2019)[23] use block-local attention that has high training speed on TPUs. This allows them to use three-dimensional data(eg. video) efficiently. Wang et al. (2020)[9] decrease the time complexity by representing 2D self-attention as 1D self-attention. Hence, this approach opens a road to apply attention to a larger region.

The first direct application of Transformers with global self-attention to full-sized im-ages is proposed in Dosovitskiy et al. (2020)[13] 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 Chen et al. (2020)[24] trains a transformer that predicts pixels while not including the structure of the image. Their model is trained on low-resolution images and learns image representations in an unsupervised manner. GAN-BERT Shin et al. (2020)[25] trains BERT for generating PET images from MRI images and their methods are easily applicable as they can scale. They chose the Next Sentence Prediction (NSP) module as their GAN discriminator. We observed that the prior research on transformers hasn'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 classification or report generation (NLG). One such example is Generating Radiology reports via Memory-driven transformer Chen and Song and Chang and Wan et al. (2020)[29]. 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.

Another example is Reinforced Transformer Xiong et al. (2019)[28]. 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.

One more example can be given as CGMVQA Ren and Zhou et al. (2020)[27], which is a medical visual question answering framework. In this paper, features extracted by ResNet121 are used by transformers.

3. Methods

We propose a Transformer-based image classification model. We first review the recently proposed Vision Transformer model Dosovitskiy et al. (2020)[13] which uses a patch-based Transformer encoder. Then, we describe how we extend their framework by introducing a Transformer decoder to perform image-to-image translation tasks.

3.1 Vision Transformer

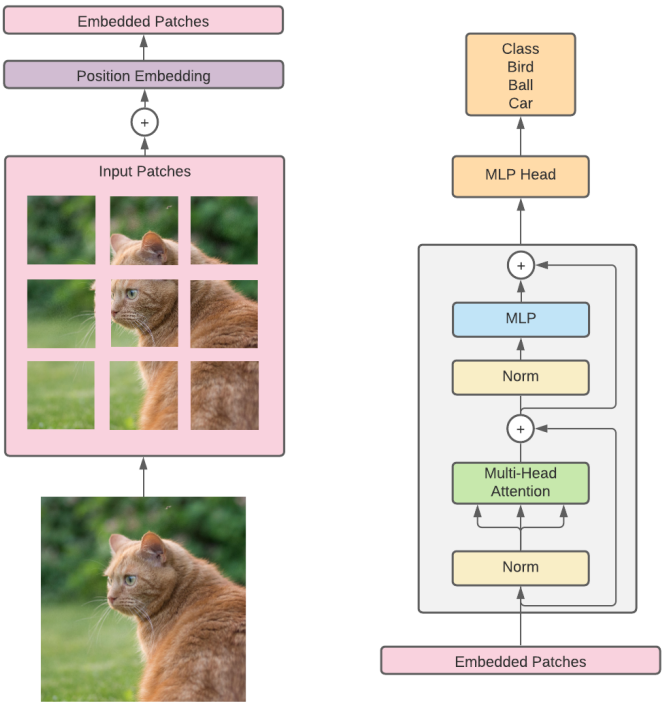

Figure 1: Vision Transformer

The original transformer framework receives input as a 1D sequence. To handle 2D sequences, Dosovitskiy et al. (2020)[13] 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 1). In other words, instead of receiving a 1D sequence of tokens, their patch-based strategy reshapes the image x ∈ RH×W×Cinto a sequence of flattened 2D patches xp ∈ RN×P2Cwhere (H, W) represents the size of the original image, (P, P) represents the size of each image patch and represents the length of the sequence for the Transformer. Position embeddings in Vision Transformers follow standard 1D position embeddings and are added to the patch embeddings to retain positional information. Vision Transformers follows the original design of Transformer encoder Vaswani et al.(2017)[1] closely, and uses multi-headed self-attention and feed-forward blocks as follows:

$$z_o = [x_{class}; x_p^1 E; \ldots; x_p^N E;] + E_{pos} \quad (1)$$

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1} \quad (2)$$

$$z_l = MLP(LN(z_l')) + z_l' \quad (3)$$

$$y = LN(z_l^0) \quad (4)$$

where MSA represents multi-headed self-attention MLP represents a multilayer perceptron that contains two layers with a GELU non-linearity. A learnable embedding is appended to the sequence of embedded patches and represents the image representation denoted with y. MSA is an extension of self-attention that projects the outputs of multiple self-attention operations run in parallel.

3.2 Strided

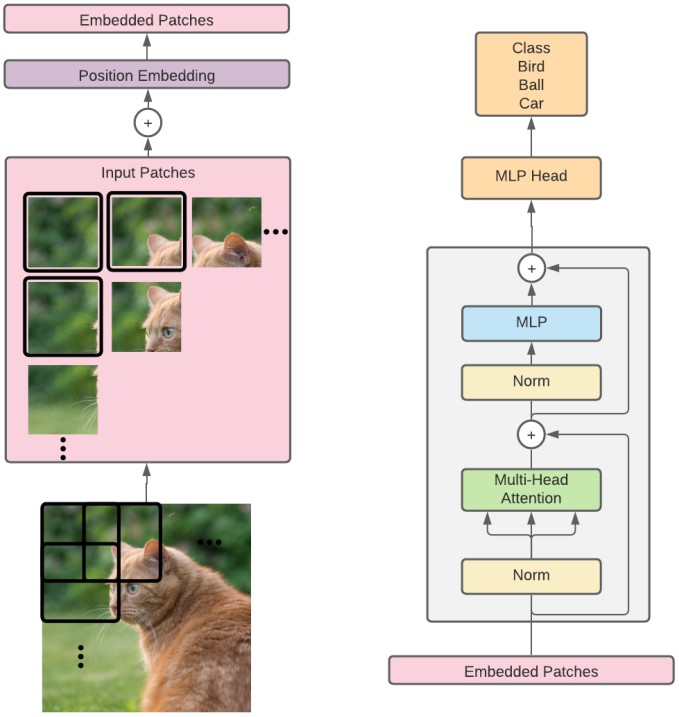

Figure 2: Strided Vision Transformer

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) Dosovitskiy et al. (2020)[13] proposes a patch-based attention mechanism. But this method loses some features that 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. 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) Dosovitskiy et al. (2020)[13]. But in our approach, it starts from N-x th pixel, so it produces more attention by using NxN patches.

3.3 Distill - pre-train

In a recent paper, Touvron and Cord et. al. (2020)[26] 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.

$$teacherLoss = teacherLoss * temperature^2 \quad (1)$$

$$loss = studentLoss * alpha + teacherLoss * (1 - alpha) \quad (2)$$

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

3.4 Applying to 3D CT Scans

In 3D CT scans visual information is stored in 3D arrays instead of 2D arrays. This makes using Vision Transformers unfeasible in a direct approach. We plan to get layers of 3D scans as in Figure 6 to get 2D vectors and then use Vision Transformers after. In the end we plan to use another neural network for combining our scores into probability scores. To do this transformation we used 3 different methods.

1. We take the average of these layers.
2. We concatenate the layers (created 2D image just as seen in Figure 6).
3. Lastly we used autoencoders to transform CT scans into 2D vectors.

3.5 Dataset

For this project, we use the Chexpert Irvin et al. (2019)[30] 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).

3.5.1 Automated Label Extraction

Chexpert Irvin et al. (2019)[30] 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. You can see an example of one patient below.

Sex	Female
Age	19
Frontal/Lateral	Frontal
AP/PA	AP
No Finding	
Enlarged Cardiomediastinum	
Cardiomegaly	
Lung Opacity	1.0
Lung Lesion	
Edema	
Consolidation	-1.0
Pneumonia	
Atelectasis	
Pneumothorax	-1.0
Pleural Effusion	-1.0
Pleural Other	
Fracture	
Support Devices	1.0

Table 1: Example Chexpert patient sample

3.5.2 Data preparation

For our purposes, we focused on 5 observations that are mentioned in related works. These are Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion. We also train our model

without differentiating between Frontal or Lateral and without differentiating PA or AP. Since we used binary classifiers for observations we needed to reduce output states from 4 to 2. For Edema and Atelectasis we changed blank and -1 labels with 0 and for other observations, we changed blank with 0 and -1 with 1.

3.5.3 3D Dataset

In the medical domain, one of the most common 3D visuals are CT scans. We used the National Institute of Health Clinical Center CT scan dataset, which includes 32,000 samples.

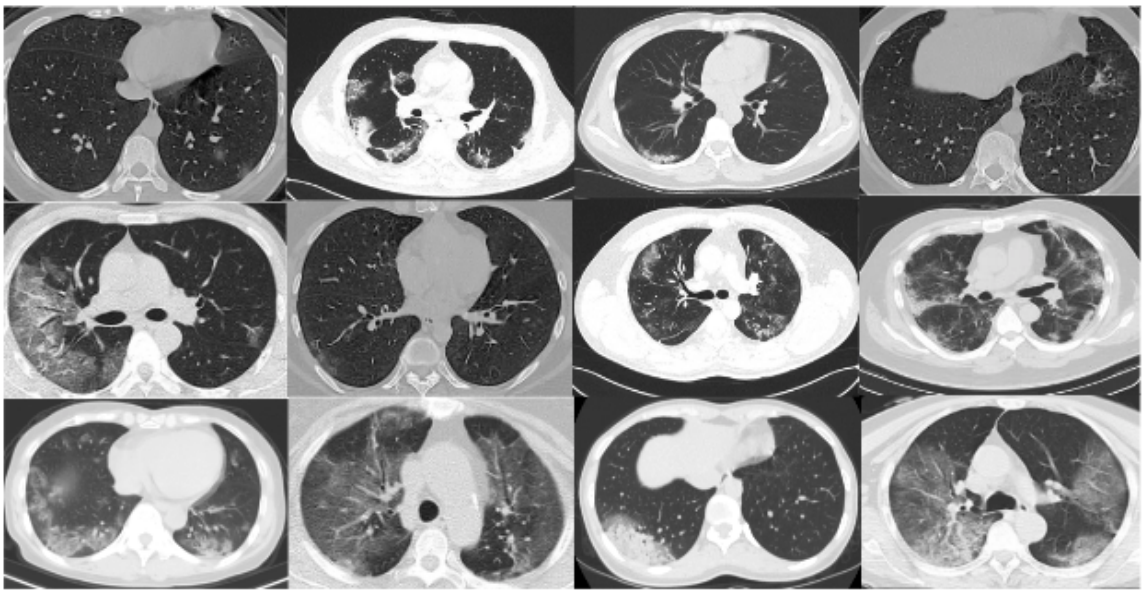

Figure 6: Layers of a CT scan concatenated in 2D

4. Results

In our experiments, we tried seven different approaches to increase our medical transformers' prediction accuracy.

4.1 ViT

Our first approach was directly using Vision Transformer (ViT) Dosovitskiy et al. (2020)[13] in the Chexpert dataset. We did the implementation using PyTorch in Python. The results we got were below the Stanford Baseline for the Chexpert dataset. Since Vision Transformer produces better results with larger datasets, this was expected.

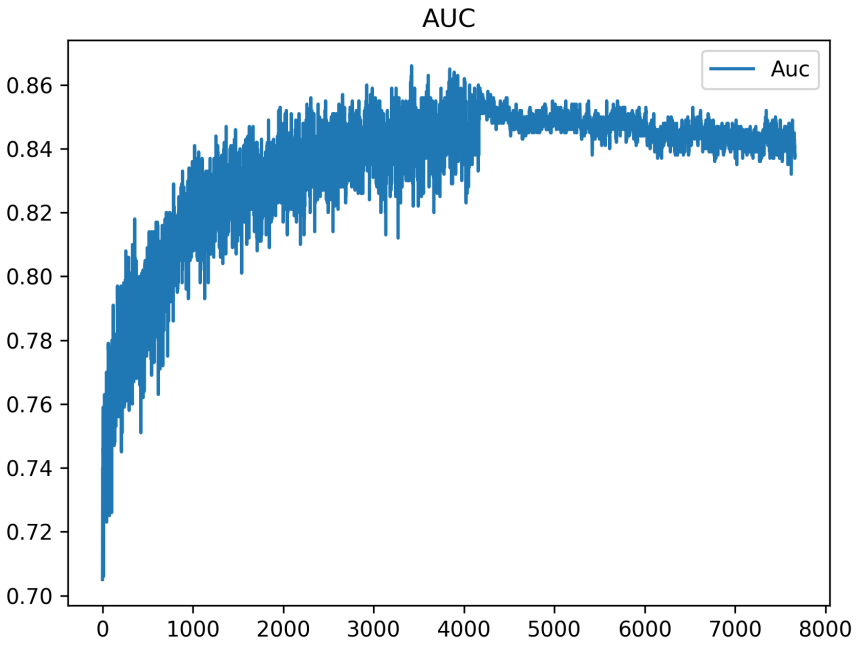

Figure 4: ViT AUC Graph

4.2 ViT + Model Ensembling

Instead of using a single model to predict 5 classes, we tried to use 5 different binary classifiers, so one binary classifier for each class. This approach was expected to produce better results than the first one, but since 5 different models consume a huge amount of memory, we needed to decrease batch size considerably. This decrease caused our model to perform poorly.

4.3 Strided ViT

Since Vision Transformer (ViT) Dosovitskiy et al. (2020)[13] follows a path-based approach in the attention layer, some information between these patches is lost. To prevent this, we followed a strided approach. Since it is a memory-heavy approach, we needed to decrease the batch size, but not as much as the previous approach. So, it could only show a little progress in terms of accuracy but proved that it has potential.

4.2 ViT + Model Ensembling

Instead of using a single model to predict 5 classes, we tried to use 5 different binary classifiers, so one binary classifier for each class. This approach was expected to produce better results than the first one, but since 5 different models consume a huge amount of memory, we needed to decrease batch size considerably. This decrease caused our model to perform poorly.

4.3 Strided ViT

Since Vision Transformer (ViT) Dosovitskiy et al. (2020)[13] follows a path-based approach in the attention layer, some information between these patches is lost. To prevent this, we followed a strided approach. Since it is a memory-heavy approach, we needed to decrease the batch size, but not as much as the previous approach. So, it could only show a little progress in terms of accuracy but proved that it has potential.

4.4 Distillable ViT

As stated above, Vision Transformer (ViT) Dosovitskiy et al. (2020)[13] needs a huge amount of input to produce better results. But Distillable Vit Touvron and Cord et. al. (2020)[26] can learn more with less data. So, since our dataset is comparatively small, we decided to try Distillable Vit Touvron and Cord et. al. (2020)[26]. For distillation, a pre-trained model is needed and we used a pre-trained DenseNet121. This approach was able to produce better results.

4.5 Distillable, Strided ViT

In this approach, we combined distillation and strided approach. As expected, it increased the accuracy of the model.

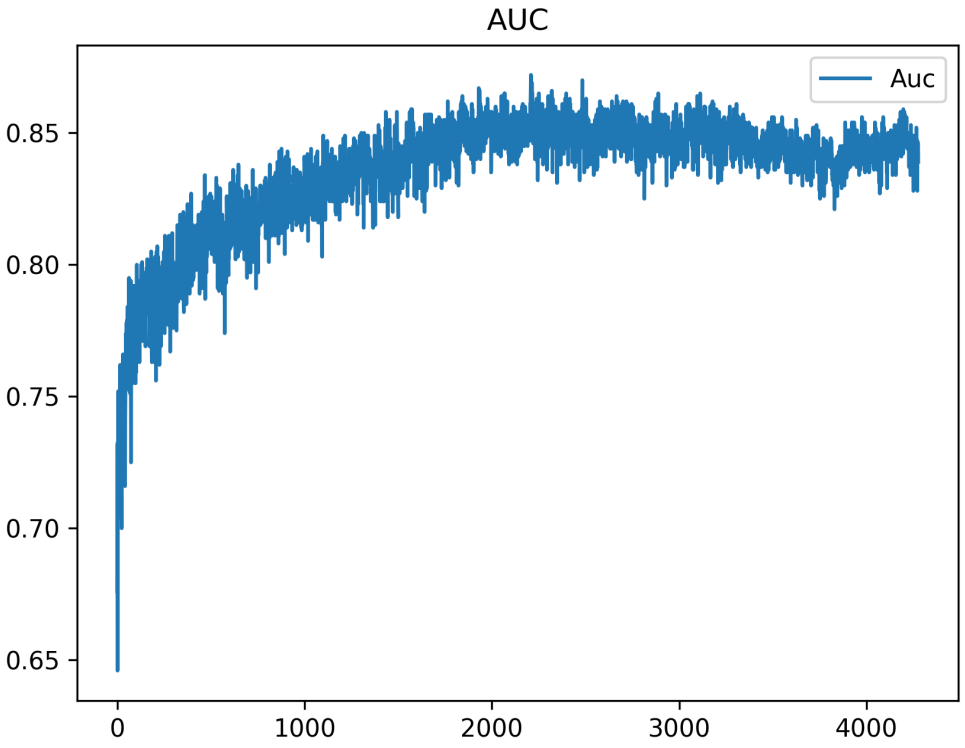

Figure 5: DS-ViT AUC Graph

	Best AUC
ViT	0.866
ME-ViT	0.753
S-ViT	0.868
D-ViT	0.872
DS-ViT	0.880

Table 2: 2D AUC Scores

4.5 Mean of Image Vectors

In our first approach for converting 3D images to 2D, we calculated the mean of the slices of image. This approach did not give significant results as expected.

4.6 Concatenating Images

As a second approach, we concatenated each slice of 3D image and created a 2D image. This approach performed better than our first approach but since the images were bigger this time, our batch size has also decreased and this affected our results in a negative manner.

4.7 3D to 2D Autoencoder

Lastly, we used a CNN Autoencoder to convert 3D images to 2D. With this approach, our results improved further.

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

Table 2: 3D AUC Scores

5. Conclusion and Discussion

5.1 Conclusions

In this work, we developed an image classification model for medical X-ray images using the Chexpert dataset. 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 close to one percent increase in mean AUC score using the strided version. This increase is especially more apparent in the distilled version of the vision transformer.

We also compared our architecture with six different variations of vision transformers that we have mentioned in the previous sections. We demonstrated that using vision transformers together with convolutional architecture achieves much better performance than vision transformers alone. Our model also managed to beat Stanford's baseline model. Thus we believe vision transformers will have potential in the medical domain.

5.2 Recommendations

We believe that increasing batch size has a positive impact on the model so we suggest increasing it as much as possible. We also think that the reason the model ensemble underperformed was due to batch size. Given enough GPU memory size, it's very likely that the model ensemble should outperform our best results.

Although we beat the best line model, our time on TRUBA was limited hence we were able to fine-tune our parameters only to a limited degree. We used handpicked values for the parameters we deemed most impactful, however we believe that if there are enough resources randomized hyperparameter search will give better results.

7. References

This is the last section of the report, before any appendices. The references should not be double-spaced, but single-spaced. For a technical report, use the CSE style.

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

[2] Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation,9(8): 1735–1780, 1997.

[3] Chung-Cheng Chiu and Colin Raffel. Monotonic chunk-wise attention. arXiv preprint arXiv: 1712.05382, 2017.

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

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

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

[7] 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.

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

[9] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-ChiehChen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. arXivpreprint arXiv: 2003.07853, 2020.

[10] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, AlexanderKu, and Dustin Tran. Image transformer. arXiv preprint arXiv: 1802.05751, 2018.

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

[12] 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.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, SylvainGelly, et al.

An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929, 2020.

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

[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, PrafullaDhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, SandhiniAgarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, AdityaRamesh, 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.

[16] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhut-dinov. Transformer-xl: Attentive language models beyond a fixed-length context. arXivpreprint arXiv: 1901.02860, 2019.

[17] 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.

[18] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451, 2020.

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

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

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

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

[23] Dirk Weissenborn, Oscar Tackstrom, and Jakob Uszkoreit. Scaling autoregressive video models. arXiv preprint arXiv:1906.02634, 2019.

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

[25] 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 trans-formers for mri to pet synthesis. arXiv preprint arXiv: 2008.04393, 2020.

[26] Hugo Touvron and Matthieu Cord and Matthijs Douze and Francisco Massa and Alexandre Sablayrolles and Herve Jegou. Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877, 2020.

[27] F. Ren and Y. Zhou, "CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering," in IEEE Access, vol. 8, pp. 50626-50636, 2020, doi: 10.1109/ACCESS.2020.2980024.

[28] Xiong, Yuxuan and Du, Bo and Yan, Pingkun. Reinforced Transformer for Medical Image Captioning. In Machine Learning in Medical Imaging, 2019.

[29] Zhihong Chen and Yan Song and Tsung-Hui Chang and Xiang Wan. Generating Radiology Reports via Memory-driven Transformer. arXiv eprint: 2010.16056, 2020

[30] Jeremy Irvin and Pranav Rajpurkar and Michael Ko and Yifan Yu and Silviana Ciurea-Ilcus and Chris Chute and Henrik Marklund and Behzad Haghgoo and Robyn Ball and Katie Shpanskaya and Jayne Seekins and David A. Mong and Safwan S. Halabi and Jesse K. Sandberg and Ricky Jones and David B. Larson and Curtis P. Langlotz and Bhavik N. Patel and Matthew P. Lungren and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv eprint: 1901.07031, 2019.