SB-SSL: Slice-Based Self-supervised Transformers for Knee Abnormality Classification from MRI

Sara Atito[1], Syed Muhammad Anwar[2(✉)], Muhammad Awais[1,3], and Josef Kittler[1]

[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK
[2] Children's National Hospital, Washington, DC, USA
sanwar@childrensnational.org
[3] Surrey Institute for People-Centred AI, Guildford, UK

Abstract. The availability of large scale data with high quality ground truth labels is a challenge when developing supervised machine learning solutions for healthcare domain. Although, the amount of digital data in clinical workflows is increasing, most of this data is distributed on clinical sites and protected to ensure patient privacy. Radiological readings and dealing with large-scale clinical data puts a significant burden on the available resources, and this is where machine learning and artificial intelligence play a pivotal role. Magnetic Resonance Imaging (MRI) for musculoskeletal (MSK) diagnosis is one example where the scans have a wealth of information, but require a significant amount of time for reading and labeling. Self-supervised learning (SSL) can be a solution for handling the lack of availability of ground truth labels, but generally requires a large amount of training data during the pretraining stage. Herein, we propose a slice-based self-supervised deep learning framework (SB-SSL), a novel slice-based paradigm for classifying abnormality using knee MRI scans. We show that for a limited number of cases (<1000), our proposed framework is capable to identify anterior cruciate ligament tear with an accuracy of 89.17% and an AUC of 0.954, outperforming state-of-the-art without usage of external data during pretraining. This demonstrates that our proposed framework is suited for SSL in the limited data regime.

Keywords: Self-supervised learning · Group masked model learning · Masked autoencoders · Knee abnormality · Transformers · MRI

1 Introduction

Knee abnormality can arise from a variety of factors including aging, physical injury, and joint disease. MRI is the standard-of-care for diagnosis of knee abnormalities [21], where the image contains a wealth of information and the scanning

S. Atito and S. M. Anwar—Contributed equally to this article.

G. Zamzmi et al. (Eds.): MILLanD 2022, LNCS 13559, pp. 86–95, 2022.
https://doi.org/10.1007/978-3-031-16760-7_9

protocols are safe from a clinical perspective. Knee MRI exams are among the most widely performed scans in MSK radiology [14]. MSK conditions arise from a variety of reasons (including sports injury and lifestyle choices) effecting adults and pediatrics. Both the amount of information within a knee MRI scan, and the number of such scans performed on a daily basis put a huge burden on the radiologist and the clinical workforce dealing with MSK related conditions and knee abnormalities In recent years, machine learning is the technology of choice in radiology for automated image analysis and abnormality identification [2]. However, the clinical translation of this technology is facing challenges such as lack of adequate annotations and training data. In particular, manual segmentation and data labeling is a labor intensive and tedious task, which is also effected by inter-rater variability. The probability of error, accounting for the day-to-day workload on radiologists, is high and this is where machine learning can benefit the most by identifying the most critical cases needing immediate attention.

In contrast to Convolutional Neural Networks (CNNs), transformer-based deep learning models have shown to perform better due to an inherent design incorporating attention and parallel computing [17]. The success of transformer based networks in the field of natural language processing (NLP) is phenomenal and became the default choice in most recent NLP applications. The recent introduction of vision transformer [10], has resulted in the translation of some of this success to vision tasks. Training self-supervised vision transformers for medical applications could alleviate some of the problems associated with acquiring high quality ground truth labels and hence, accelerate the research in computer aided diagnosis. However, such networks require a large training data. Therefore, in Computer Vision (CV) problems, the default practice is to use a pretrained model on a large supervised data like ImageNet-1K, before fine tuning for a specific downstream task with limited data [4].

Recently, self-supervised pretraining of deep neural networks without using any labels has outperformed supervised pretraining in CV [3,5]. This phenomenal shift in CV is less investigated in medical image analysis domain. We argue that recent SSL approaches are ideally suited for medical image analysis, since medical data are an order of magnitude smaller than natural images due to several reasons, including privacy concerns, expensive annotation, rarity of certain diseases, etc. Hence for medical applications, SSL can lead the way for a wider adoption of such techniques in domains where labels are not available or are difficult to acquire [1]. Therefore, the purpose of this study is to investigate: 1) is ImageNet-1K pretraining needed for medical imaging? 2) can we perform self-supervised pretraining on a small medical data and outperform large scale out of distribution supervised pretraining? If successful this will form the basis for SSL for medical imaging in limited data regimes. Towards this, we propose a slice-based self-supervised deep learning framework (SB-SSL) for abnormality classification using knee MRI, where our main contributions are:

– We propose a novel slice based self-supervised transformer model (SB-SSL) for knee abnormality classification using magnetic resonance imaging data.

– The model is pretrained from scratch on limited data without labels and fine tuned for the downstream knee abnormality classification task with state-of-the-art performance.
– Our experimental results show that, when trained using the group masked model learning (GMML) paradigm, SSL can be successfully applied for medical image analysis with limited data/label.

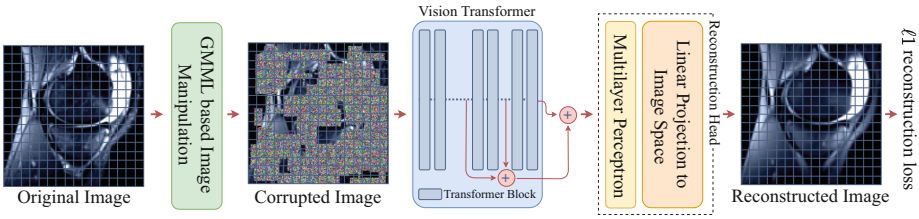


Fig. 1. Proposed self-supervised learning approach.

2 Related Works

In [6], a deep learning based method was presented for the detection of abnormalities in knee MRI. The publicly available MRNet data was presented, along with an AlexNet [16] based model for classifying abnormalities, meniscal tear, and anterior cruciate ligament (ACL) tear. This was among the first approaches where deep learning was applied to this task, and since then has been used in multiple studies to further improve the classification performance [11,13,20,24].

A CNN based self supervised training paradigm was developed, where solving the jigsaw puzzle was used as the pre-text task [20]. In the downstream task, ACL tear was classified with an accuracy of 76.62% and an area under the curve (AUC) of 0.848 using the sagittal plane. In [24], efficiently-layered network (ELNet) was proposed where the model reduced the number of parameters compared to AlexNet, and utilized individual slice views for classification of meniscus (coronal) and ACL (axial) tears. An accuracy of 0.904 with an AUC of 0.960 was achieved in detecting the ACL tear. This performance was improved by adding a feature pyramid network and pyramidal detail pooling to ELNet [11]. An AUC of 0.976 and an accuracy of 0.886 was achieved in ACL tear classification task. However, both these methods are based on supervised training. Meniscus tears were identified using a deep learning model and compared with manual evaluation [13]. An accuracy of 95.8% was achieved for an internal validation set, however the model was not evaluated on any of the publicly available data.

In general, it should be noted that for methods that report higher performance, training is based on the availability of ground truth labels. Whereas for self supervised training, which could alleviate this burden, the model performance drops. We propose, for the first time, a transformer based self-supervised framework for knee abnormality classification using MRI. Our innovative training paradigm use self-supervised training and shows that such a framework can be effectively used even when the size of training data is relatively small.

3 Methodology

In this work, we introduce a general slice-based self-supervised vision transformer for knee MRI medical records. The system diagram of the proposed approach is shown in Fig. 1. Transformers [25] have shown great success in various NLP and CV tasks [3–5,7–9,22,27] and are the basis of our proposed framework.

3.1 Vision Transformer

Vision transformer [10] receives, as input, a feature map from the output of a convolutional block/layer with K kernels of size $p \times p$ and stride $p \times p$. The convolutional block takes an input image $\mathbf{x} \in \mathcal{R}^{C \times H \times W}$ and converts it to feature maps of size $\sqrt{n} \times \sqrt{n} \times K$, where C, H, and W are the number of channels, height, and width, of the input image, $(p \times p)$ is the patch size, and n is the number of patches, i.e., $n = \frac{H}{p} \times \frac{W}{p}$. Learnable position embeddings are added to the patch embeddings as an input to the transformer encoder to retain the relative spatial relation between the patches.

The transformer encoder consists of L consecutive Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. The MSA block is defined by h self-attention heads, where each head outputs a sequence of size $n \times d$. The self attention mechanism is based on a trainable triplet (query, key, and value). Each query vector in $\mathbf{Q} \in \mathcal{R}^{n \times d}$ for a given head is matched against a set of key vectors $\mathbf{K} \in \mathcal{R}^{n \times d}$, scaled by the square root of d to have more stable gradients as the dot product of q and k tend to grow large in magnitude, resulting in vanishing gradients and a slowdown of learning. After applying softmax, the output is then multiplied by a set of values $\mathbf{V} \in \mathcal{R}^{n \times d}$. Thus, the output of the self-attention block is the weighted sum of $\mathbf{V}$ as shown in Eq. 1. The output sequences across heads are then concatenated into $n \times (d \times h)$, and projected by a linear layer to a $n \times K$ sequence. The MLP block consists of two point-wise convolution layers with GeLU [12] non-linearity.

$$\text{SelfAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}. \tag{1}$$

3.2 Self-supervised Pretraining

We leverage the strength of the transformers and train it as an autoencoder with a light decoder employing GMML [4,5]. Starting with the vanilla transformer autoencoder, the model is pretrained as an autoencoder to reconstruct the input image, i.e., $D(E(\mathbf{x})) = \mathbf{x}$, where $\mathbf{x}$ is the input image, E is the encoder which is vision transformer in our case, and D is a light reconstruction decoder. Due to the strength of transformers, it is expected that the model will perfectly reconstruct the input image after a few training epochs. Indeed, this is attributed to the fact that without a proper choice of constraints, autoencoders are capable of learning identity mapping, i.e., memorizing the input without learning any useful discriminative features.

To promote the learning of context and learn better semantic representations of the input images from the transformer-based autoencoder, we apply several transformations to local patches of the image. The aim is to recover these masked local parts at the output of the light decoder. In doing so, especially with a high percentage of corruption (up to 70%), the model implicitly learns the semantic concepts in the image and the underline structure of the data in order to be able to recover the image back. Image in-painting is a simple but effective pre-text task for self-supervision, which proceeds by training a network to predict arbitrary transformed regions based on the context.

The objective of image reconstruction is to restore the original image from the corrupted image. For this task, we use the $\ell 1$-loss between the reconstructed image and the original image in an end-to-end self-supervised trainable system as shown in Eq. 2. Although, $\ell 2$-loss generally converges faster than $\ell 1$-loss, it is prone to over-smooth the edges for image restoration [26]. Therefore, $\ell 1$-loss is more commonly used for image-to-image processing.

$$\mathcal{L}(\mathbf{W}) = \sum_{k}^{b} \left(\sum_{i}^{H} \sum_{j}^{W} \mathbb{1}_{[\mathbf{M}_{i,j}^{k}=1]} |\mathbf{x}_{i,j}^{k} - \bar{\mathbf{x}}_{i,j}^{k}| \right), \tag{2}$$

where $\mathbf{W}$ denotes the parameters to be learned during training, b is the batch size, $\mathbf{M}$ is a binary mask with 1 indicating the manipulated pixels, and $\bar{\mathbf{x}}$ is the reconstructed image. To further improve the performance of the autoencoder, we introduced skip connections from several intermediate transformer blocks to the decoder. These additional connections can directly send the feature maps from the earlier layers of the transformers to the decoder which helps to use fine-grained details learned in the early layers to construct the image. Besides, skip connections in general make the loss landscape smoother which leads to faster convergence. Further, the reconstructed image $\bar{\mathbf{x}}$ is obtained by averaging the output features from the intermediate blocks from the transformer encoder ($E(.)$) and feeding the output to a light decoder ($D(.)$) represented mathematically as $\bar{\mathbf{x}} = D\left(\sum_{i \in \mathcal{B}} E_i(\hat{\mathbf{x}})\right)$, where $E_i(.)$ is the output features from block i and $\mathcal{B}$ is a pre-defined index set of transformer blocks that are included in the decoding process. Herein, we set $\mathcal{B}$ to $\{6, 8, 10, 12\}$.

As for the decoder, unlike CNN-based autoencoders which require expensive decoders consisting of convolutional and transposed convolution layers, the decoder in the transformer autoencoder can be implemented using a light decoder design. Specifically, our decoder consisted of two point-wise convolutional layers with GeLU non-linearity and a transposed convolutional layer to return back to the image space. Since the backbone, i.e., vision transformer, and the light decoder are isotropic, some of the transformer blocks may act as decoder and hence, heavy and computationally expensive type of decoders are not required.

4 Experimental Results

To demonstrate the effectiveness of our proposed self-supervised vision transformer on medical images, we employed the MRNet dataset [6]. The dataset

consists of 1,370 knee MRI records, split into a training set of 1,130 records of 1,088 patients and a validation set of 120 records of 111 patients. Each MRI is labeled according to the presence/absence of meniscus tear, ACL tear, or any other abnormality in the knee. In this work, we tackled the ACL tear identification problem using the Sagittal plane. The dataset is highly imbalanced with only 208 MRIs representing ACL tear.

4.1 Implementation Details

In our work, we employed the ViT Small (ViT-S) variant of the transformer [23] with 256×256 input image size. For optimization of the transformer parameters during self-supervised pre-training, we used the Adam optimizer [19] with a momentum of 0.9. The weight decay follows a cosine schedule [18] from 0.04 to 0.4, and a base learning rate of $5e^{-4}$. All models were pre-trained employing 4 Nvidia Tesla V100 32 GB GPU cards with 64 batch size per GPU.

Simple data augmentation techniques were applied like random cropping, random horizontal flipping, random Gaussian blurring, and random adjusting of the sharpness, contrast, saturation, and the hue of the image. The augmented image was further corrupted by randomly replacing patches from the image with zeros, with a replacement rate of up to 70% of the image pixels.

For fine-tuning, we drop the light decoder and fine tune the pre-trained model by passing the volume, slice by slice, to the transformer encoder. The outputs of the class tokens corresponding to each slice are then concatenated to obtain $y \in \mathcal{R}^{f \times K}$, where f is the number of slices. After that, the features y are fed to a fully connected layer with K nodes followed by GeLU non-linearity, followed by a linear layer with 2 nodes corresponding to the presence/absence of the ACL tear. As the dataset is highly imbalanced, we used oversampling on the training set to balance the dataset. Specifically, we over-sample the minority class, i.e., presence of ACL tear, to match the number of the majority class. Finally, we applied the same optimization parameters and data augmentations used for the self-supervised training.

Further, we employed ensemble learning [15]. Generally, neural networks have high variance due to the stochastic training approach that makes them sensitive to the nature of the training data. The models may find a different set of weights each time they are trained, which in turn may produce different predictions. To mitigate this issue, for each experiment, we trained 5 models with different weight initialization and combined the predictions from these models. Not only this approach reduced the variance of the predictions, but also resulted in predictions that were better than any single model.

4.2 Results

It is well known that transformers are data-hungry which make them hard to train, mostly, due to the lack of the typical inductive bias of convolution operations. Consequently, the common protocol for self-supervised learning with transformers is to pretrain the model on a large scale dataset, such as ImageNet or

even larger datasets. We compare our proposed approach with the state-of-the-art SSL methods when the pretraining and the fine-tuning are performed only on the MRNet dataset. Table 1 shows that our method outperforms the state-of-the-art with a large margin with an improvement of 12.6% top-1 validation accuracy on the ACL tears classification task employing the sagittal plane. Most importantly, without using any external data, our proposed approach outperforms the competitors that are pre-trained with ImageNet-1K marking a milestone for the medical domain. The receiver operating characteristic (ROC) curve for three transformer variants, ViT-Tiny, ViT-Small, and ViT-Base are shown in Fig. 2, where ViT-T performs the best.

Table 1. Comparison with SOTA on ACL tears classification employing sagittal plane.

Method	Backbone	# params	ACL Tear (Sagittal plane)	
			Accuracy (%)	AUC
Training using only the given dataset				
Random Init	CNN	77M	71.67	0.754
Random Init	ViT-S	21M	70.00	0.721
[20]	CNN	77M	76.62	0.848
[20] + noise	CNN	77M	75.83	0.817
SB-SSL (Ours)	ViT-T	5M	85.83	0.952
SB-SSL (Ours)	ViT-S	21M	88.33	0.954
SB-SSL (Ours)	ViT-B	86M	89.17	0.954
Transfer learning from ImageNet-1K dataset				
MRNet [6]	AlexNet	61M	86.63	0.963

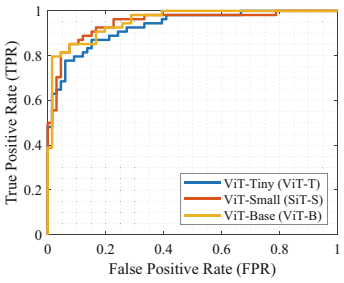
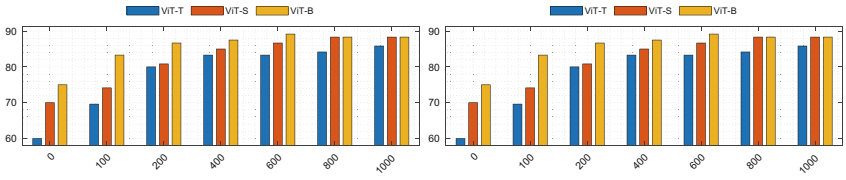


Fig. 2. ROC curves of the classification of ACL tears employing different vision transformer architectures.

4.3 Ablation Studies

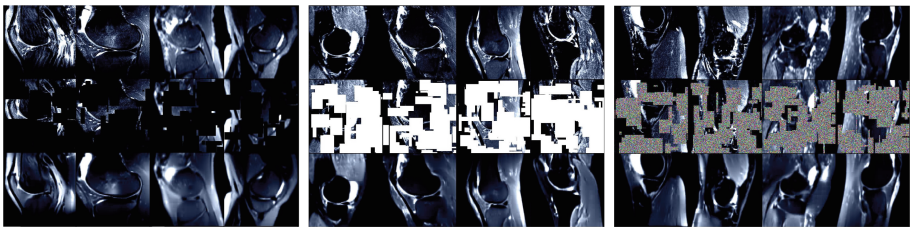
In this section, we investigate the effect of different recipes of the proposed approach, such as the effect of longer pretraining, the size of the model, and the type of image corruption during the pretraining stage. Further, we show the interpretability of the system by visualizing the attention of the trained models.

Effects of Longer Pretraining and Model Size. In Fig. 3, we show the performance of the proposed approach when pretrained for longer duration across different vision transformer architectures. The x-axis represents the number of self-supervised pretraining epochs, with zero indicating that the model was not pretrained, i.e., training from scratch. From the reported results, it is evident that the training from random initialization has produced a lower accuracy as the amount of data available is insufficient to train the transformer. The results significantly improved when the models were pretrained without any external data by 25.8%, 18.3%, and 13.3% employing ViT-T, ViT-S, and ViT-B, respectively, compared to training from scratch. Another observation is that pre-training the self-supervised for longer and employing bigger transformer architectures contribute positively to the performance of the proposed approach.



(a) Accuracy (b) AUC

Fig. 3. Top-1 validation accuracies and AUC of the MRNet validation set across different vision transformer architectures. The x-axis represents number of epochs used for pretraining.



(a) Zeros (b) Ones (c) Noise

Fig. 4. Samples of different types of corruption. The rows represent the original images, corrupted images, and the reconstructed images after the pretraining stage, respectively.

The Effects of Different Types of Corruption: We first investigate the effect of training a vanilla transformer autoencoder, where the model is pretrained as an autoencoder to reconstruct the input image. As expected, after finetuning, the performance was similar to the performance of the model trained from scratch. Following, we investigate the effect of applying different types of image inpainting including: random masking by replacing a group of connected patches from the image with zeros, ones, or noise. Samples of the different types of corruption are shown in Fig. 4 along with the reconstructed images after the pretraining stage. The performance when pretraining the models with different types of corruption is on par, with noise being marginally better than others.

Attention Visualization. To verify that the model is learning pertinent features, in Fig. 5, we provide visualizations of the self-attention corresponding to the class token of the 10^{th} layer of the vision transformer. To generate the attention for an image, we compute the normalized average over the self-attention heads to obtain a 16×16 tokens. The tokens are then mapped to a color scheme, up-sampled to 256×256 pixels, and overlaid with the original input image. For visualization, we selected the mid slice of randomly selected MRI volumes from the MRNet validation set. We observe that the attention is clearly focusing on the area of interest, corresponding to the main part of the MRI slice on which the detection of ACL tears is performed.

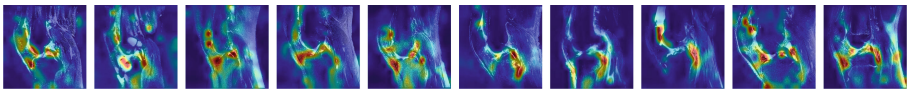


Fig. 5. Self-attention visualizations from the ViT-S model finetuned on the ACL tears task employing the sagittal plane.

5 Conclusion

We proposed a novel framework SB-SSL, pre-trained in a self-supervised manner for knee abnormality classification. We established a new benchmark in SSL for MRI data, where pretraining on a large supervised data was not required. The state-of-the-art performance, with an accuracy of 89.17% in ACL tear classification, shows that our proposed method can be employed in MR image classification even when the data are limited and ground truth labels are not available.

References

1. Anwar, S.M., et al.: Semi-supervised deep learning for multi-tissue segmentation from multi-contrast MRI. J. Signal Process. Syst. 1–14 (2020)
2. Anwar, S.M., et al.: Medical image analysis using convolutional neural networks: a review. J. Med. Syst. **42**(11), 1–13 (2018)
3. Atito, S., Awais, M., Farooq, A., Feng, Z., Kittler, J.: MC-SSL0.0: towards multi-concept self-supervised learning. arXiv preprint arXiv:2111.15340 (2021)
4. Atito, S., Awais, M., Kittler, J.: SiT: Self-supervised vision transformer. arXiv preprint arXiv:2104.03602 (2021)
5. Atito, S., Awais, M., Kittler, J.: GMML is all you need. arXiv preprint arXiv:2205.14986 (2022)
6. Bien, N., et al.: Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. PLoS Med. **15**(11), e1002699 (2018)
7. Brown, T.B., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
8. Chen, Z., et al.: Masked image modeling advances 3d medical image analysis. arXiv preprint arXiv:2204.11716 (2022)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Dunnhofer, M., Martinel, N., Micheloni, C.: Improving MRI-based knee disorder diagnosis with pyramidal feature details. In: Medical Imaging with Deep Learning, pp. 131–147. PMLR (2021)
12. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUS). arXiv preprint arXiv:1606.08415 (2016)
13. Hung, T.N.K., et al.: Automatic detection of meniscus tears using backbone convolutional neural networks on knee MRI. J. Magn. Reson. Imaging (2022)

14. Irmakci, I., Anwar, S.M., Torigian, D.A., Bagci, U.: Deep learning for musculoskeletal image analysis. In: 2019 53rd Asilomar Conference on Signals, Systems, and Computers, pp. 1481–1485. IEEE (2019)
15. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**, 1–9 (2012)
17. Liu, Y., et al.: A survey of visual transformers. arXiv preprint arXiv:2111.06091 (2021)
18. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
19. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in Adam. arXiv:abs/1711.05101 (2017)
20. Manna, S., Bhattacharya, S., Pal, U.: Self-supervised representation learning for detection of ACL tear injury in knee MR videos. Pattern Recogn. Lett. **154**, 37–43 (2022)
21. Nacey, N.C., Geeslin, M.G., Miller, G.W., Pierce, J.L.: Magnetic resonance imaging of the knee: an overview and update of conventional and state of the art imaging. J. Magn. Reson. Imaging **45**(5), 1257–1275 (2017)
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
23. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877 (2020)
24. Tsai, C.H., Kiryati, N., Konen, E., Eshed, I., Mayer, A.: Knee injury detection using MRI with efficiently-layered network (ELNET). In: Medical Imaging with Deep Learning, pp. 784–794. PMLR (2020)
25. Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
26. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imaging **3**(1), 47–57 (2016)
27. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image analysis. arXiv preprint arXiv:2203.05573 (2022)