Thesis Template

Tobias Wrigstad

April 17, 2024

Abstract

Replace this with the actual abstract. Obviously. Suspendisse luctus leo et porta mattis. In semper, nisi et suscipit iaculis, leo urna laoreet lacus, ut laoreet lorem tellus eget dui. Vestibulum eu auctor nisi. Morbi pharetra euismod velit ac mattis. Maecenas tempor vitae augue ut aliquam. Nunc auctor, nibh at imperdiet finibus, ex leo semper lacus, ac vehicula quam nisl condimentum leo.

Contents

Chapter 1

Introduction

X-ray microtomography uses X-rays to create cross-sections of real objects, which can then be utilised to create a virtual model without causing any damage to the original object. It can be used for both industrial computed tomography and medical imaging. Microtomography, often known as micro-CT, has the capability of rapidly achieving high resolution. Because micro-CT makes it possible for scientists and medical professionals to investigate the morphology, physiology, and pathology of many tissues and organs, it is useful and important in the field of medical imaging. Furthermore, the creation of innovative implants and biomaterials, as well as the assessment of their performance and biocompatibility, can be facilitated by micro-CT. While micro-CT is a fantastic tool for evaluating medical imaging, there are certain obstacles that must be addressed. The prefix "micro" indicates that the cross-sections' pixel sizes are within the micrometre range. The virtual model has a high resolution and is extremely massive. For example, a 1 cubic centimetre volume representing a mouse tumour requires more than 50 GB of storage. This enormous amount of data presents a major challenge to processing procedures. In order to properly compress the original data and learn some useful representations, individuals are turning to computer vision for assistance these days.

Computer vision, on the other hand, has a rich history that encompasses decades of research into enabling computers to understand visual stimuli meaningfully. Computer vision has grown in importance in scientific disciplines as deep learning matures and powerful computing resources such as GPUs become more widely available. Using large datasets, computer vision models may learn a variety of pattern recognition skills and accomplish complex tasks such as classification and segmentation. The fundamental deep learning technology used in these tasks is the convolutional neural network (CNN). CNN is a standard supervised learning method, hence data labels are required. This can be especially difficult for medical data because labeling medical data needs experts and could be very expensive. As a result, self-supervised learning and transfer learning are popular techniques in which a model is trained on a large unrelated corpus (such as ImageNet) before being fine-tuned on a dataset of interest. However, natural image data has distinct features from medical image data, and pre-trained models trained on natural images do not produce satisfactory results for medical images, like the result shown below in Figure 1.1. To address the challenge of obtaining large medical imaging datasets, Zhou et al. (2023) suggested an approach that involves pre-training models on the same dataset as the downstream dataset, yielding satisfactory results.

The combination of computer vision with medical imaging can benefit the mod-

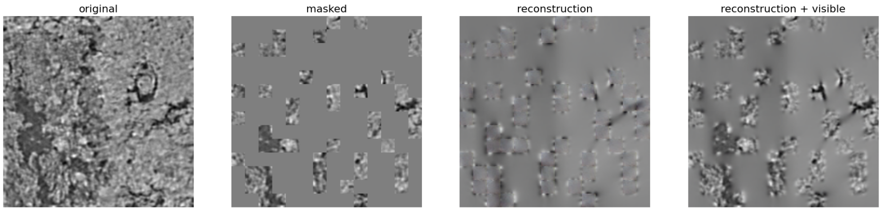

Figure 1.1: Micro-CT reconstruction with model pre-trained on ImageNet.

ern medical imaging sector in a variety of ways, including diagnosis, prediction of future outcomes, pathological segmentation from organs to cells, disease monitoring, and so on. Based on the presented challenges of managing micro-CT data, we intend to use computer vision methods, notably autoencoders, to develop an effective approach for compressing and encoding data, as well as to exploit our newly discovered latent representations for further research.

Chapter 2

Background

Machine learning is now widely used across many industries, and it has become a powerful tool that may help people in many ways. For example, machine learning techniques can be used to quickly and accurately analyze large amounts of data, while human specialists can continuously evaluate and improve the results. In medical imaging, machine learning has become more and more popular since it can extract important information that helps with precise diagnosis (Giger 2018). Even though supervised learning has dominated for years, labeling large amounts of training data is becoming prohibitively expensive and time-consuming. As a result, self-supervised learning has become the preferred method for situations involving a moderate amount of annotated data. The following parts introduce the most essential and common self-supervised learning techniques and their applications in medical imaging.

2.1 Self-supervised learning

In contrast to supervised learning, self-supervised learning employs labels created from the data itself (Liu et al. 2023), negating the need for manual labeling. This possibility leads to more data-efficient models by enabling models to learn from vast amounts of easily accessible data without the requirement for manual annotation. Based on the differences in model structure, self-supervision can be divided into three categories: generative, contrastive, and predictive.

Generative Learning The generative learning technique arises from the concept that if the model can generate or reconstruct one image, it must have learned something from it. One of the most popular generative models is the autoencoders, which was first introduced by Ballard (1987) for pre-training artificial neural networks. Autoencoder learns efficient representations by training the network to ignore signal noise. The autoencoder consists of an encoder that compresses the input into a latent representation, and a decoder that reconstructs the input from the latent representation. The goal of the autoencoder is to minimize the reconstruction error, which measures how well the decoder can reproduce the original input. Bidirectional encoder representations from transformers (BERT) (Devlin et al. 2018) is one of the state-of-the-art autoencoder models. Two tasks are used to pre-train BERT (Devlin et al. 2018) on enormous amounts of unlabeled data. The first challenge is masked language modeling, which entails forecasting tokens that have been masked after a certain fraction of input tokens have been hidden. For the model to comprehend sentence relationships, the second challenge is the next sentence prediction. The BERT (Devlin et al. 2018) design is straightforward and effective in trial data. Another popular type of generative model is the auto-regressive model, which predicts the next token in a sequence based on the previous tokens. GPT (Brown et al. 2020, Radford et al. 2018, 2019) is one of the most prominent examples of auto-regressive models, which also uses a transformer-based neural network to encode and decode the input sequence. Auto-regression models have also been employed in computer vision, such as PixelRNN (Van Den Oord, Kalchbrenner & Kavukcuoglu 2016) and PixelCNN (Van den Oord, Kalchbrenner, Espeholt, Vinyals, Graves et al. 2016), which use auto-regression methods to model images pixel by pixel. Inspired by the great success of BERT, He et al. (2022) proposed a masked autoencoders (MAE) method for computer vision. MAE uses vision transformers as the backbone, masking random patches of the input image and reconstructing the missing pixels to learn the general pattern of the input dataset.

Contrastive Learning "Learn to compare" is the goal of contrastive learning (Gutmann & Hyvärinen 2010). Using a similarity metric to group similar data closer together and different samples farther apart is the fundamental concept of contrastive learning (Jaiswal et al. 2020). Given that the data lacks labels, one potential solution to the label problem is the use of pseudo labels, as introduced by Deep Cluster (Caron et al. 2018). This technique uses clustering to produce pseudo labels, which are then used to ask a discriminator to predict the labels of images. However, this clustering method is time-consuming and performs poorly when compared to later instance discrimination-based algorithms. Instance discrimination methods boost performance by introducing efficient data augmentation strategies. Contrastive multiview coding (CMC) (Tian et al. 2020) proposes using multiple views of an image as positive samples and another as negative sample. They discovered that the more views they learned, the better the resultant representation reflected the underlying scene semantics. Similarly, SimCLR (Chen et al. 2020) employs various data augmentation methods, demonstrating the importance of the composition of data augmentation in learning accurate representations. SimCLR (Chen et al. 2020) outperforms CMC (Tian et al. 2020) in managing negative samples on a large scale. MoCo (He et al. 2020) employs momentum contrast to increase the amount of negative samples. It eliminates the typical end-to-end framework and instead uses two encoders (query and key), which reduces the fluctuation of loss convergence in the initial period.

Predictive Learning Predicting future or missing information is a frequent strategy for sequential data analysis. Text prediction has been a focus of research in natural language processing for decades. Using a pretext task is the process of creating labels so that supervised approaches can be used in unsupervised situations to train models. According to Mikolov et al. (2013), high-quality word vectors may be trained while reducing computational complexity dramatically. Word2Vec (Mikolov et al. 2013) uses "center word prediction" as a pretext task. It predicts the missing word between a sequence of words, allowing the model to acquire word representations that may then be used to train models for subsequent tasks. Other language models also use various types of pretext challenges to develop their models' comprehensive sentences. For instance, GPT (Brown et al. 2020, Radford et al. 2018, 2019) and BERT (Devlin et al. 2018) utilize next word prediction and word prediction from both sides, respectively.

2.2 Vision transformer

Transformers (Vaswani et al. 2017) are neural network architectures capable of processing sequential data. It was introduced in natural language processing and quickly became the method of choice. However, the application of transformers in computer vision was not very successful until the introduction of Vision Transformers (Dosovitskiy et al. 2020). Transformer's fundamental improvement is the replacement of all recurrent layers with pure attention and fully connected blocks. However, attention does not include a sense of location inside a sequence, therefore, we must add position information to each token. There are many choices of positional encodings, learned and fixed (Gehring et al. 2017). Transformer employs a smart positional encoding method, representing each position with a sinusoid rather than a direct index vector. Sinusoidal position (Vaswani et al. 2017) provides a high value for close tokens that gently decays when one looks at tokens that are further away, and avoids being large in magnitude for extended sequences. The mathematical definition of sinusoidal position is given below:

$$P_{(p,2m)} = \sin\left(\frac{p}{10000^{\frac{2m}{d_{\text{model}}}}}\right)$$
$$P_{(p,2m+1)} = \cos\left(\frac{p}{10000^{\frac{2m}{d_{\text{model}}}}}\right)$$

where p is the position, m is the dimension, and d_{model} is the dimension of the output embedding space. Since P_{p+k} can be expressed as a linear function of P_p for any fixed offset k, this enables the model to learn to attend by relative locations.

2.3 Self-supervised methods in medical image analysis

Medical imaging analysis's main goal is to extract useful data that supports precise diagnosis (Anwar et al. 2018) and facilitates treatment planning and follow-up for clinics. Classification, detection and localization, segmentation, and registration are the four main tasks in medical imaging that sprang out of computer vision tasks (Altaf et al. 2019). However, a significant issue is the lack of high-quality annotated medical imaging datasets. Finding medical images to curate takes a lot of time and money. This is where self-supervised learning comes into play. Self-supervised learning offers the ability to pre-train models so that they capture the general appearance of medical images using the available unlabeled data.

However, self-supervised learning on medical imagery presents various obstacles. Medical images, unlike natural images, appear to be similar. The distinction between healthy and ill patients is based on small visual signals that are highly localized and often difficult to detect (Huang et al. 2023). As a result, some essential phases in self-supervised learning are potentially problematic.

The introduction of several kinds of self-supervised learning used in medical imaging analysis will follow.

Generative Generative approaches appear to be the least popular method in medical imaging. Chen et al. (2019) developed and validated a novel self-supervised learning technique based on context restoration for classification, localization, and segmentation. They demonstrate that this context restoration strategy outperforms existing pre-training models and provides considerable performance improvements over the baseline. Zhou et al. (2023) proposed self-pre-training using the MAE (He et al. 2022) approach. They pre-trained a ViT on the target data's training set rather than another dataset, and discovered that MAE self-pre-training significantly improves a variety of medical image tasks such as classification, CT multi-organ segmentation, and MRI brain tumor segmentation.

Constrastive Contrastive learning is currently the most popular method in medical imaging. Since some standard augmentation methods may be inadequate for medical imaging, instead of building positive pairings with different enhanced versions of the same image, one can improve positive sampling according to the similarity of a patient's clinical information (Huang et al. 2023). For instance, images from the same patient (Azizi et al. 2021) and images from patients of similar ages (Dufumier et al. 2021). Some researchers attempted to employ pretext tasks based on natural images as well. Several studies (Sowrirajan et al. 2021, Vu et al. 2021, Chen et al. 2021, Sriram et al. 2021) used the MoCo (He et al. 2020) framework to create pre-trained models. According to Sowrirajan et al. (2021), not all of the augmentation procedures used in the MoCo article are suitable for gray-scale images. Sowrirajan et al. (2021) solely used random partial rotation and horizontal flipping in their experiments. Sriram et al. (2021) purely used MoCo (He et al. 2020) and trained on non-COVID chest X-ray images. They also used a continuous positional embedding module to extract representations from a collection of time-indexed radiographs. In addition, SimCLR (Chen et al. 2020) is also widely used in medical imaging. Chaitanya et al. (2020) provided two significant improvements by developing domain-specific and problem-specific knowledge simultaneously. The method yields substantial improvements compared to other self-supervision and semi-supervised learning techniques.

Predictive Predictive approaches are currently gaining popularity. According to Zhang et al. (2017), a typical 3D CT or MR volume contains rich context information, which may be easily indexed in 2D slices. The order of slices can be employed in a pretext task to train a predictive self-supervised learning model. This method takes advantage of the particular qualities of medical images and achieves promising results with minimal use of human annotations. Nguyen et al. (2020) proposed a spatial awareness method that can learn semantic and spatial representations from volumetric medical images. The method consists of two steps: spatial awareness learning and semantic feature learning. In the first step, the method randomly crops four patches from an input image and predicts their relative positions using a convolutional neural network (CNN). This task forces the network to learn the spatial structure and layout of the image. In the second step, the method uses the same CNN to extract semantic features from the whole image and compares them with the features of another image using a contrastive loss function. This task encourages the network to learn discriminative and robust features that can distinguish different images. The results showed that the proposed method can improve the performance of organ segmentation and intracranial hemorrhage detection tasks by using the learned features as initialization or regularization for the supervised models.

Chapter 3

Methodology

The micro-CT data used in this project offers exceptional resolution and includes information from both low and high magnification. For a medical task, both types of information may be relevant for certain activities, so we want to keep multiscale data. However, because this data is not manageable in size, we wish to first divide the entire image into manageable-sized chunks. We opted to try out the masked autoencoder (MAE)(He et al. 2022) technique in the project.

The decision to attend MAE was motivated by two factors. First, patch-based processing allows the model to collect implicit information at low magnification. Masking certain parts encourages the model to learn the full-scale information as well. This coincides with our goal of preserving multiscale data. Second, the structure of MAE is quite straightforward and easy to develop. Furthermore, the structure of the model considerably reduces the computational resources it requires.

3.1 Masked autoencoder

Masked autoencoder is a simple autoencoding approach that reconstructs the original input signal given its partial observation. MAE consists of an encoder and a decoder. The encoder part of the MAE transforms the visible image patches into a latent representation of M dimensions. This representation is later used as input for the decoder so that the decoder can reconstruct the whole input image. The structure of MAE is illustrated in Figure 3.1.

MAE relies on vision transformers (ViT) (Dosovitskiy et al. 2020) as its foundation. First, we partition the image into non-overlapping patches. As in a standard ViT, we project these patches into vectors and add a fixed sinusoidal position (Vaswani et al. 2017) to each one. In this method, we provide our model with relative location information, allowing it to retain knowledge of patch order. Figure 3.2 shows an example of embedded position information. Then we sample a subset of patches and mask (i.e., remove) the remainder. The masking strategy is simple: generate random numbers between $(0, 1)$ using a uniform distribution, take these numbers as the index of tokens, sort them in ascending order, and then remove a specified percentage of the input tokens.

By randomly masking a high percentage of patches (in our example 70%), our encoder acts on a small part of the overall set, significantly lowering the processing resources required. This allows us to train enormously large encoders with a fraction of the computational resources and memory. Random masking also plays an important function in eliminating redundancy. Unlike texts, images have low semantic

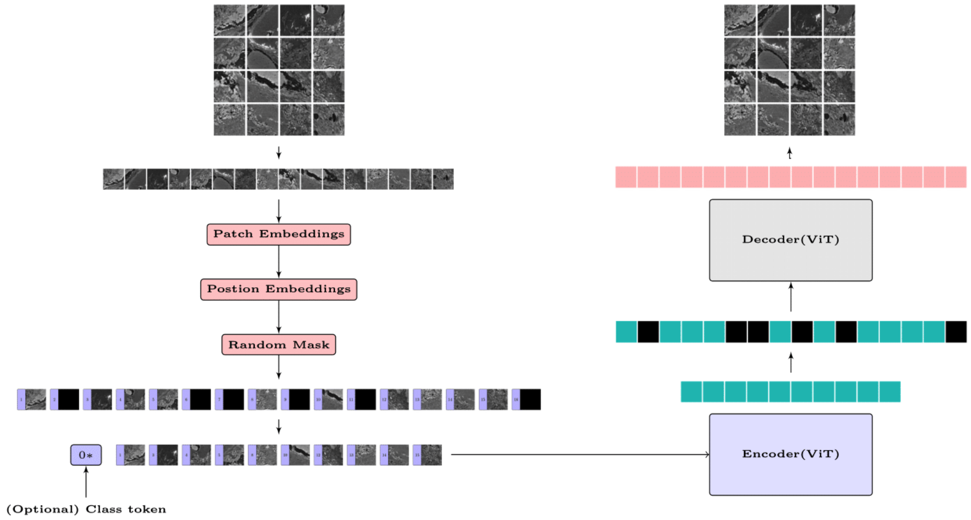

Figure 3.1: MAE architecture illustration.

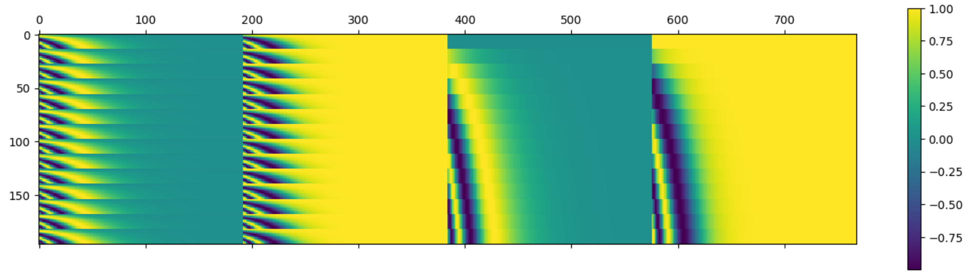

Figure 3.2: Position embedding visualization for input data.

information, thus we must manually design a difficult assignment to encourage the model to learn high-level semantic information, instead of simply reconstructing the missing pixels from neighboring patches.

The encoder is simply a standard ViT (Dosovitskiy et al. 2020) applied exclusively to visible patches. Following the ViT (Dosovitskiy et al. 2020) steps, the encoder embeds patches using a linear projection with extra positional embeddings, and then processes the resulting set using a sequence of Transformer blocks. As previously stated, our encoder only operates on a small subset of the total input patches, all masked tokens are eliminated, and no mask tokens are used. A lightweight decoder handles the entire set.

Figure 3.1 illustrates how the decoder handles all tokens. Each masked token is a common, learnt vector that signals the presence of a missing patch that must be predicted. By adding position embeddings to all tokens in this whole set, masked tokens may determine where they belong in the image. This masked set, along with the encoded visible tokens, passes through the decoder Transformer blocks to produce the reconstructed input image. Moreover, the decoder is independent of the encoder, so the architecture can be flexibly designed. In the experiment, we used a shallower and narrower decoder to process the whole set, which reduces the pre-training time.

The MAE decoder reconstructs the input by predicting the pixel values of each masked patch. The decoder's output contains a vector of pixel values representing a patch. The decoder output is reshaped to provide a reconstructed image. We employ mean squared error:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.1}$$

to compute the loss between the reconstructed and original images in pixel space. We calculate the loss just on masked patches.

TODO

- Augmentation of the input images
- Clustering method

Chapter 4

Experiments

4.1 Set up

We do self-supervised pre-training on the micro-CT dataset. Then we perform clustering on the same dataset but with unseen data (different parts of the slices).

4.2 Model training

4.3 Main Properties

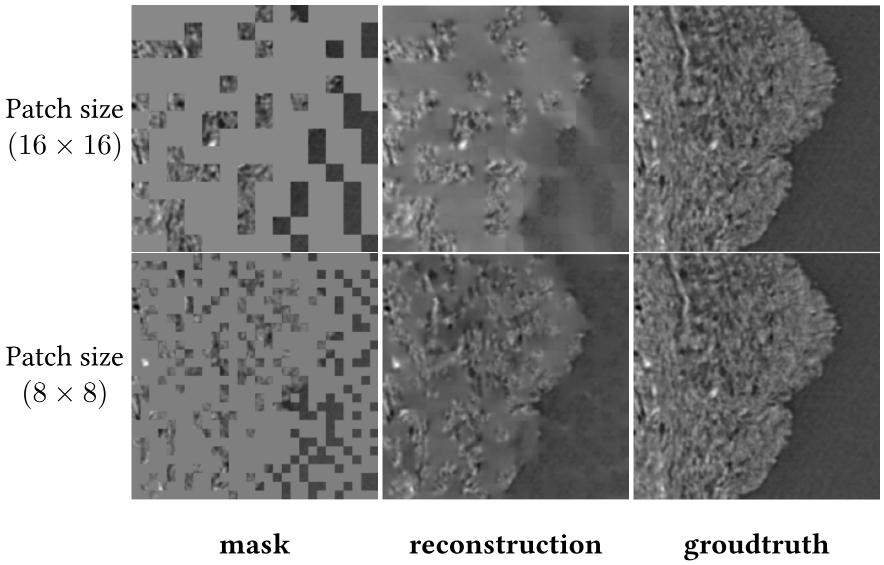

Figure 4.2: Results of reconstructed images with different patch size.

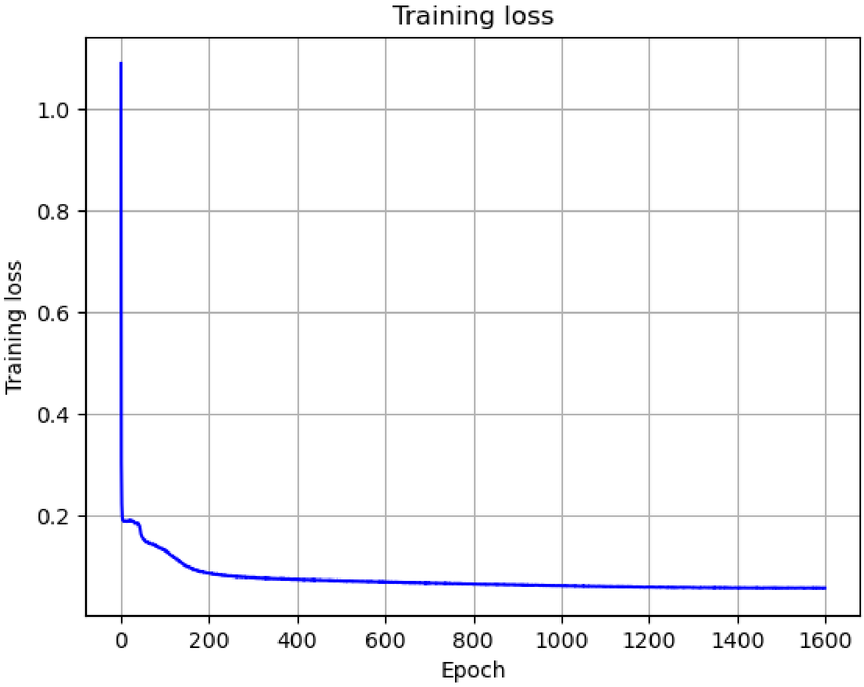

Figure 4.1: Training loss per epoch.

Chapter 5

Implementation

Chapter 6

Discussion

Chapter 7

Conclusion

Bibliography

Altaf, F., Islam, S. M., Akhtar, N. & Janjua, N. K. (2019), 'Going deep in medical image analysis: concepts, methods, challenges, and future directions', *IEEE Access* **7**, 99540–99572.

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M. & Khan, M. K. (2018), 'Medical image analysis using convolutional neural networks: a review', *Journal of medical systems* **42**, 1–13.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T. et al. (2021), 'Big self-supervised models advance medical image classification', *Proceedings of the IEEE/CVF international conference on computer vision* pp. 3478–3488.

Ballard, D. H. (1987), 'Modular learning in neural networks', *Proceedings of the sixth National Conference on artificial intelligence-volume 1* pp. 279–284.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020), 'Language models are few-shot learners', *Advances in neural information processing systems* **33**, 1877–1901.

Caron, M., Bojanowski, P., Joulin, A. & Douze, M. (2018), 'Deep clustering for unsupervised learning of visual features', *Proceedings of the European conference on computer vision (ECCV)* pp. 132–149.

Chaitanya, K., Erdil, E., Karani, N. & Konukoglu, E. (2020), 'Contrastive learning of global and local features for medical image segmentation with limited annotations', *Advances in neural information processing systems* **33**, 12546–12558.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M. & Rueckert, D. (2019), 'Self-supervised learning for medical image analysis using image context restoration', *Medical image analysis* **58**, 101539.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020), 'A simple framework for contrastive learning of visual representations', *International conference on machine learning* pp. 1597–1607.

Chen, X., Yao, L., Zhou, T., Dong, J. & Zhang, Y. (2021), 'Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images', *Pattern recognition* **113**, 107826.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020), 'An image is worth 16x16 words: Transformers for image recognition at scale', *arXiv preprint arXiv:2010.11929* .

Dufumier, B., Gori, P., Victor, J., Grigis, A., Wessa, M., Brambilla, P., Favre, P., Polosan, M., McDonald, C., Piguet, C. M. et al. (2021), 'Contrastive learning with continuous proxy meta-data for 3d mri classification', *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24* pp. 58–68.

Gehring, J., Auli, M., Grangier, D., Yarats, D. & Dauphin, Y. N. (2017), 'Convolutional sequence to sequence learning', *International conference on machine learning* pp. 1243–1252.

Giger, M. L. (2018), 'Machine learning in medical imaging', *Journal of the American College of Radiology* **15**(3), 512–520.

Gutmann, M. & Hyvärinen, A. (2010), 'Noise-contrastive estimation: A new estimation principle for unnormalized statistical models', *Proceedings of the thirteenth international conference on artificial intelligence and statistics* pp. 297–304.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. (2022), 'Masked autoencoders are scalable vision learners', *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. (2020), 'Momentum contrast for unsupervised visual representation learning', *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 9729–9738.

Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S. & Chaudhari, A. S. (2023), 'Self-supervised learning for medical image classification: a systematic review and implementation guidelines', *npj Digital Medicine* **6**(1), 1–15.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D. & Makedon, F. (2020), 'A survey on contrastive self-supervised learning', *Technologies* **9**(1), 2.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J. & Tang, J. (2023), 'Self-supervised learning: Generative or contrastive', *IEEE Transactions on Knowledge and Data Engineering* **35**(1), 857–876.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Nguyen, X.-B., Lee, G. S., Kim, S. H. & Yang, H. J. (2020), 'Self-supervised learning based on spatial awareness for medical image analysis', *IEEE Access* **8**, 162973–162981.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018), 'Improving language understanding by generative pre-training'.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019), 'Language models are unsupervised multitask learners', *OpenAI blog* **1**(8), 9.

Sowrirajan, H., Yang, J., Ng, A. Y. & Rajpurkar, P. (2021), 'Moco pretraining improves representation and transferability of chest x-ray models', *Medical Imaging with Deep Learning* pp. 728–744.

Sriram, A., Muckley, M., Sinha, K., Shamout, F., Pineau, J., Geras, K. J., Azour, L., Aphinyanaphongs, Y., Yakubova, N. & Moore, W. (2021), 'Covid-19 prognosis via self-supervised representation learning and multi-image prediction', *arXiv preprint arXiv:2101.04909* .

Tian, Y., Krishnan, D. & Isola, P. (2020), 'Contrastive multiview coding', *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16* pp. 776–794.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A. et al. (2016), 'Conditional image generation with pixelcnn decoders', *Advances in neural information processing systems* **29**.

Van Den Oord, A., Kalchbrenner, N. & Kavukcuoglu, K. (2016), 'Pixel recurrent neural networks', *International conference on machine learning* pp. 1747–1756.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.

Vu, Y. N. T., Wang, R., Balachandar, N., Liu, C., Ng, A. Y. & Rajpurkar, P. (2021), 'Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation', *Machine Learning for Healthcare Conference* pp. 755–769.

Zhang, P., Wang, F. & Zheng, Y. (2017), 'Self supervised deep representation learning for fine-grained body part recognition', *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)* pp. 578–582.

Zhou, L., Liu, H., Bae, J., He, J., Samaras, D. & Prasanna, P. (2023), 'Self pre-training with masked autoencoders for medical image classification and segmentation', *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* pp. 1–6.