Optimizing the Interpretability of Contrastive Learning Models: An Analysis of Augmentation Strategies

by

Amy Saranchuk
Supervisor: Michael Guerzhoy

April 2024

Abstract

Contrastive self-supervised learning is a promising approach in fields where labelled training data is limited, such as medical imaging. A model is trained to distinguish between augmented versions of the same image and completely different images. This allows it to learn representations of unlabeled data, which are then used to fine-tune a model for a downstream task, such as image segmentation. This work starts by fine-tuning a contrastive learning model on three medical imaging datasets. Using rotation as the augmentation strategy, it then generates saliency maps using the SmoothGrad technique for pairs of input images at incremental rotation angles θ for each dataset. It is hypothesized that the greater the similarity between the highlighted region of the saliency map and the ground truth segmentation mask, the more interpretable the results will be. This measure of interpretability is compared against increasing rotation angles, which reveals a distinct periodic pattern, unique to each dataset. This pattern, with a periodicity of 90°, suggests that model interpretability varies with rotation, achieving peaks or troughs at 90° intervals, depending on the dataset. While it was initially hypothesized that this may be due to the model's reliance on Histogram of Oriented-Gradients (HOG)-like features, subsequent experiments did not support this theory. The discovery of optimal augmentation angles in this research has the potential to guide the preparation of datasets for training highly interpretable models in the field of medical imaging.

Acknowledgement

I extend my deepest gratitude to my supervisor, Professor Guerzhoy, for his unwavering guidance and invaluable feedback throughout this journey. His support was instrumental in the completion of this work. I would also like to thank my parents and sister for their ongoing encouragement and support throughout my academic endeavors. Lastly, a special note of thanks to my cat for her companionship and moral support, which brightened many moments along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, self-supervised learning (SSL) has shown immense promise in reducing the need for labeled data when training machine learning models. This is achieved with an auxiliary pretext task, which aims to capture the semantic features of the data into learned representations that are subsequently applied to a downstream task of interest [9][29][11].

Within SSL, a particularly promising method is called contrastive learning. The idea behind this method is that various augmentations (such as flipping, rotating, or blurring) of the same image hold the same semantic information [30][10]. A contrastive learning model is trained to bring semantically-similar data closer together in an embedding space and push dissimilar data apart [21][11][10]. These learned representations can then be used for a downstream task.

One domain in which contrastive learning has shown significant promise is the field of medical imaging. As the number of patient images generated with respect to the number of radiologists increases, it becomes more critical to automate the process of segmenting and classifying anomalies in MRI scans [3]. However, with the lack of labeled data available in the medical field, traditional machine learning paradigms are not sufficient, hence the promising application of contrastive learning. Within the past few years, there has been a surge of research geared toward the use of SSL for medical imaging and numerous techniques have been proposed.

However, one aspect of the model that is crucial in its adoption in industry is its explainability and interpretability. While explainable AI techniques for contrastive learning remain largely unexplored, a few approaches have been proposed including averaged transforms, pairwise occlusion, and Interaction-CAM [23].

In this research, we will be applying various interpretability techniques on several medical imaging datasets. Rather than altering the model architecture, our goal is to explore and identify any patterns or insights that could inform the augmentation strategy when training a contrastive learning model to achieve the most interpretable results. In order to accomplish this, the following objectives will be pursued:

1. Train both a supervised and self-supervised model for medical image segmentation and compare their

results.

2. Apply various interpretability techniques using the self-supervised model and select the most interpretable approach.

3. Apply the chosen technique to input images that have been augmented over a continuous range.

Using the results obtained from objective 3, we will conclude whether certain augmentation strategies result in higher interpretability.

This thesis is organized into eight chapters to methodically outline and discuss the results of each of the aforementioned objectives. Chapter 2 provides the necessary background for the material discussed in this work. Chapter 3 describes each dataset used for this research. Chapter 4 contains the methodology and results for training both the supervised and self-supervised models. Chapter 5 selects an interpretability technique and applies it to input images which have been rotated at incrementing angles θ. Chapter 6 attempts to explain the patterns observed in the previous chapter. Chapter 7 provides a discussion summarizing the key claims from this work, and Chapter 8 concludes the study with potential future research paths.

Chapter 2

Background

2.1 Self-Supervised Learning

2.1.1 What is self-supervised learning?

The concept of self-supervised learning (SSL) initially emerged in the field of robotics, in which the technique involved automatically labeling data using the relationship between various sensor signals [30]. In SSL, the data supervises itself during training instead of using labels as an indication of its performance by creating artificial supervisory signals from unlabeled data. It allows the network to leverage unlabeled data by learning meaningful representations without manual annotations [21][5]. An important component of SSL algorithms is the encoder, which is responsible for mapping the input samples to a latent embedding space [11].

SSL is used across a variety of tasks, including object detection in computer vision, image comprehension, and image segmentation [21]. Most approaches to learning visual representations without human supervision can be categorized into one of two classes. The first set of approaches is generative, where the model learns to generate information in the input space. A drawback to these approaches is their computational expense. The other type is discriminative, where the model is trained to perform an auxiliary pretext task where both the labels and inputs are derived from an unlabeled dataset [4]. In this review we will focus on discriminative approaches because of their reduced computational cost.

SSL is comprised of two stages: the pretext task and the downstream task. The overall idea behind self-supervised learning is to first pre-train the model using the pretext task and then further fine-tune it for a specific task. The pretext task is used for pre-training and guides the model toward learning intermediate representations of the data. The downstream task is the task which we wish to solve, and to which we wish to transfer knowledge from the pretext task [25].

Pretext tasks are unrelated to the downstream task of interest and are used to generate pseudo labels from data without human annotation. They force the network to learn information about the data such as where edges, colours, and shapes appear in an image [9]. Designing an appropriate pretext task requires

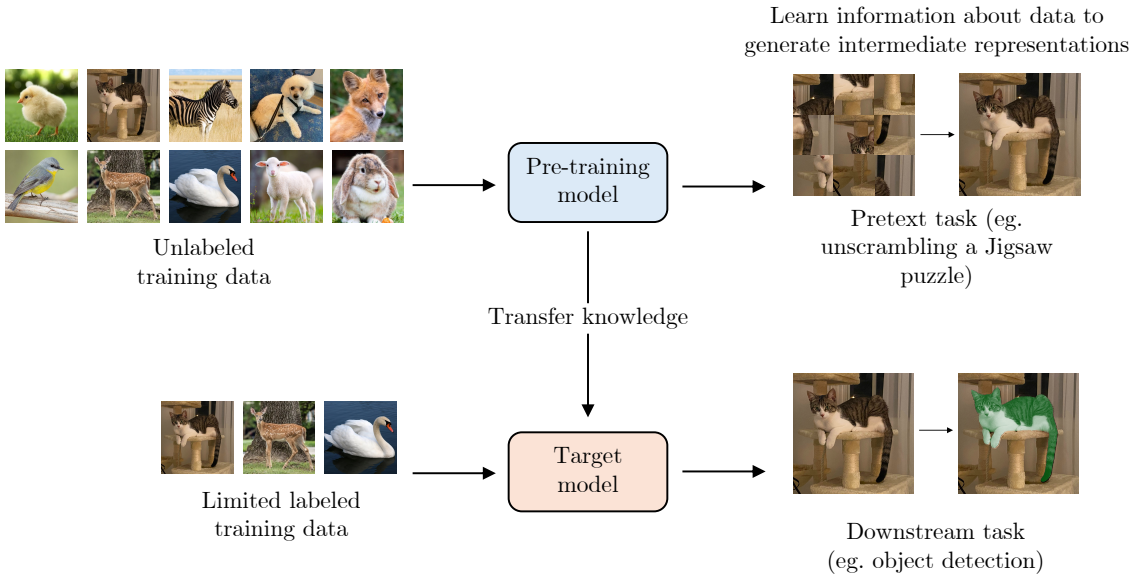


Figure 2.1: Structure of self-supervised learning algorithms.

knowledge in the domain of the downstream task (for example, medical imaging). It has been found that a smaller gap between the nature of the pretext and downstream tasks results in better performance on the downstream task itself [21]. Some common pretext tasks include recolouring an image, predicting the rotation angle, and unscrambling a Jigsaw puzzle [30].

The downstream task is the primary task of interest (for example, image classification or segmentation). The downstream task can be accomplished through either fine-tuning the pre-text model, or using a linear classifier, both of which are elaborated in Section 2.1.7 [21].

A popular discriminative pretext approach is called contrastive learning. An early proposal of this method occurred in 2013 by Mikolov et al. in the field of natural language processing [11][15]. Since then, it has been adapted to other domains and according to [4], has "revolutionized" the field of computer vision through learning meaningful representations that can be applied to a variety of vision tasks from unlabeled data.

The general idea behind contrastive learning is that various transformations of the same image still hold the same semantic information [30][10]. The goal is to bring data which is semantically-similar closer together in an embedding space, while pushing dissimilar data apart [21][11][10].

The following outlines the contrastive learning algorithm. First, each sample image from the training set is augmented. The image along with its augmentation are considered a positive pair, and the image along with every other image in the dataset are considered negative pairs. See Fig. 2.2 for an example. The model is then trained so that it can learn to distinguish between positive and negative pairs by minimizing the distance between positive pairs, and maximizing the distance between negative pairs. By doing so, the

model learns quality representations of each image, the knowledge of which is subsequently transferred to the downstream task [11]. Some loss functions used for contrastive learning include pairwise loss, triplet loss, N-pair loss, and most commonly, InfoNCE [9]. Most contrastive learning models use some variant of the ResNet model due to its balance between its large size and excellent learning capability [11].

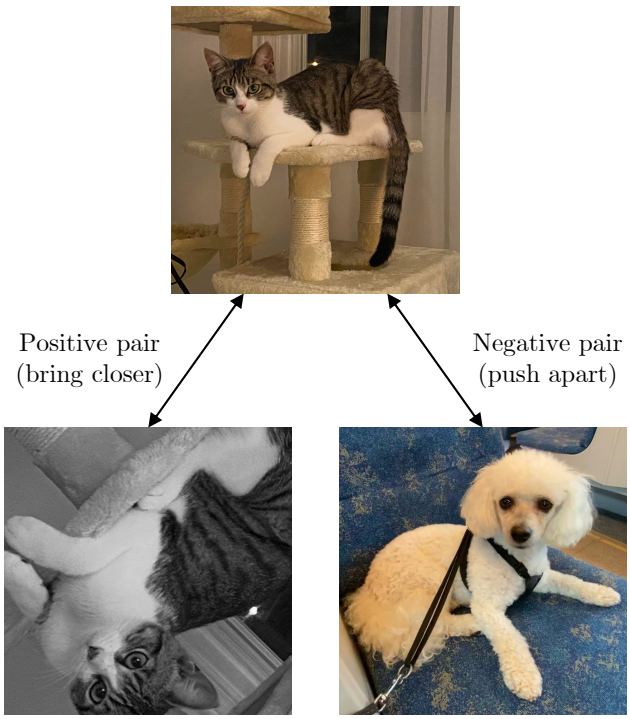


Figure 2.2: Example positive and negative pairs. At the top is the original image. Bottom left is the augmented image. Bottom right is a different image in the dataset.

Some common data augmentations for creating positive pairs include cropping, flipping, rotating, and adding noise to the image. However, like pretext tasks, not all data augmentations are equally effective. It's crucial to choose augmentations that preserve the semantic information of the data. For example, cropping may crop out integral components of the image [9].

While the neural network used for contrastive learning is more complex as opposed to non-contrastive learning, contrastive-based discriminative approaches have shown great promise as of late, and have been able to achieve state-of-the-art results [21]. For example, the proposed contrastive framework in [4] (which will be explained in more detail in Section 2.1.3) was able to match or surpass the accuracy of supervised learning for a variety of natural image classification datasets. While the accuracy of SSL may not significantly surpass that of supervised learning, its primary advantage lies in the reduced requirement for labeled data.

2.1.2 Motivation behind SSL

Labelling data by hand is a highly time-consuming and laborious process [21]. SSL was created to combat the high cost of labelling data and as an attempt to bring us closer to "embedding human cognition into machines" [25].

Unlike supervised learning, which is designed to perform a specific task, SSL gives us the ability to develop more generic AI systems while making use of unlabelled data [21]. An SSL model can adapt to various downstream tasks through fine-tuning without introducing bias from labels [23][25]. This is especially important in fields like medical imaging, where one pre-trained model can be used for a variety of downstream tasks without having to curate a large labelled dataset each time [10].

Contrastive learning is particularly effective because it improves the generalization and robustness of the network as compared to other SSL methods by distinguishing between positive and negative samples [9]. According to [29], contrastive learning is the most successful SSL technique and has been able to achieve results close to supervised learning.

2.1.3 Frameworks

The field of SSL, and particularly contrastive learning, has been rapidly growing over the past few years. There have been numerous frameworks proposed to improve vanilla contrastive learning, with some of the most popular being SimCLR and MoCo [10].

In 2020 Chen et al. proposed a novel contrastive-based SSL method called SimCLR, standing for a Simple Framework for Contrastive Learning of Visual Representations. This method outperformed supervised models on the ImageNet dataset using 100 times fewer labels [10][4]. SimCLR is a contrastive learning algorithm without the use of a memory bank. The framework is composed of the following components. First, like regular contrastive learning, each image in the dataset is augmented to create positive pairs. The augmentations used include random cropping, random colour distortions, and random Gaussian blur. A neural network base encoder (using a ResNet-50 model) then extracts representation vectors from the augmented data samples. Next, a small multilayer perceptron (MLP) projection head maps the representations to a 128-dimension latent space where the contrastive loss function is then applied. Some key differences between vanilla contrastive learning and SimCLR are the latter's use of a large batch size to allow for more negative samples in each batch, and the use of the normalized temperature-scaled cross-entropy (NT-Xent) loss. SimCLR produces representations which achieve state-of-the-art results when used for 2D natural image classification [4].

While the original paper for SimCLR focuses on 2D images, [2] proposes a 3D SSL method based on SimCLR. They claim that using a Bayesian neural network with Monte Carlo dropout during the inference phase can improve the performance on downstream tasks [2].

One drawback to SimCLR is its high computational expense due to its requirement for a very large batch

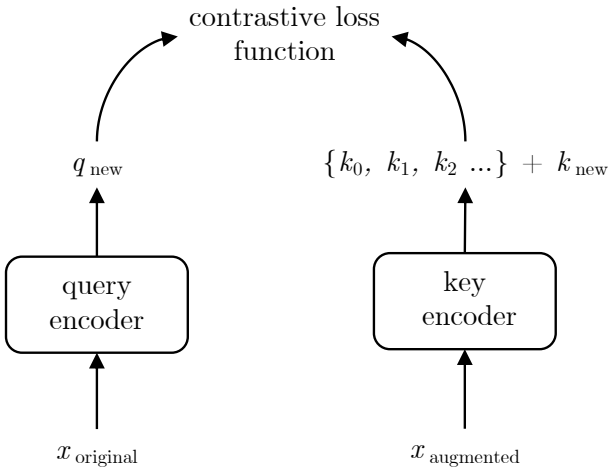


Figure 2.3: Depiction of MoCo architecture: Original image is passed through the query encoder and augmented image is passed through the key encoder. The output from the key encoder k_{new} is added to the dictionary $\{k_0, k_1, k_2, \ldots\}$ and passed into the loss function with the q_{new}. Adapted from [8].

size in order to perform well. Another method called Momentum Contrast (MoCo), also proposed in 2020, reduces this requirement by introducing a momentum-encoded queue to retain negative samples [10]. In a setup like SimCLR, the number of negative samples available to the contrastive loss is limited by the size of the batch. MoCo overcomes this by using a dynamically-sized queue of data samples, which provides a larger number of negative samples while maintaining a smaller batch size for training.

To explain this in more detail, an image from the training set is augmented into two versions (for example, a flipped version and a blurred version). Each version is passed through a separate encoder: the query encoder and key encoder. Both output vector representations of their respective images. The output of the key encoder is added to a dynamically-sized dictionary which stores negative samples with respect to the output of the query encoder. The contrastive loss is then calculated between the query encoder's output and the vectors in the dictionary. See Fig. 2.3 for a visual depiction of this. Over time, this process allows the model to learn effective representations of the images in the dataset [8].

2.1.4 Challenges in SSL

Some of the biggest challenges for SSL include the large amount of data required, its computational efficiency, and choosing an appropriate pretext task [25]. In fact, the success of an SSL model depends heavily on the quality of the pretext task chosen. An important consideration to make when choosing a pretext task is the inductive bias it may introduce into the model. It's crucial to make sure this bias is applicable to the domain of the problem. For example in contrastive learning, some data augmentations may introduce a bias into the model. This may be beneficial in encouraging invariance in transformed images, but may also be harmful depending on the data and downstream task of interest. [11] provides a helpful example: if the

desired downstream task is to detect the orientation of a photograph, using rotation as a data augmentation may significantly downgrade the model's performance. In addition, a poorly designed pretext task may cause the algorithm to find trivial solutions that it can use as a shortcut to learning the representations [5]. Choosing an appropriate pretext task is no easy feat, and although numerous methods have been proposed, it is still an active area of research [11].

2.1.5 Comparison to other methods

SSL can be seen as similar to transfer learning in that it learns representations from a pretext task and then uses these transformations for a downstream task. In transfer learning, a model is pre-trained on a large, labelled dataset. The weights of this model are then frozen and an adaptation layer is added on top. A new network is then fine-tuned using a smaller dataset with limited labels for the downstream task. The key difference here is that SSL does not require the large labelled dataset for the pretext task. In fact, the same data may be used for both the pretext and downstream tasks. This allows SSL to overcome some of transfer learning's weaknesses. For example, transfer learning may struggle in learning suitable representations due to the visual and semantic differences between the data used in the pretext task versus the data used in the downstream task. This drawback is especially prevalent in medical imaging where the nature of the images is so different from that of natural images [5]. By using the same data for both the pretext and target tasks, SSL has been found to be more effective than transfer learning [30].

In comparison to generative approaches as discussed earlier, generative adversarial networks (GANs) can have difficulty converging. Discriminative approaches like contrastive learning do not have this same shortcoming [11].

2.1.6 SSL in medical imaging

SSL has achieved remarkable success in numerous areas, including healthcare [21]. Medical imaging technology has become an essential part of modern medical practice and has improved the efficiency of diagnosing and planning treatment, as evident from the consistent rate of growth of medical imaging utilization in the industry [10]. However, assembling large-scale medical imaging datasets requires domain knowledge, is costly, and is time-consuming, limiting the efficacy of medical imaging models [10][30]. In a field like healthcare where there is such a variety in the tasks that can be automated, it is practically impossible to develop a large-scale, accurate, specialized dataset for every process we wish to automate [10]. Furthermore, according to [10] studies have shown that the average radiologist needs to interpret a medical image every 3-4 seconds to meet the demand. Consequently, delays in diagnosis and human error are essentially unavoidable. This has led to a critical need for automation in the process of analyzing medical images to improve both the efficiency and accuracy of diagnoses [10]. With the vast amount of unlabelled medical data routinely created during clinical practice and research, SSL is a promising approach [5][30].

Areas containing abnormalities in medical images (such as tumours) are called regions of interest (ROI), and it is these ROI that we wish to identity with machine learning [29]. Due to the lack of structure in medical images, using SSL is challenging and we cannot simply apply the pretext tasks designed for natural images [29][21]. Rather, we must design particular solutions for medical images specifically.

Many supervised medical imaging models struggle to generalize well to a variety of tasks. Since they are trained for a very specific task, it is difficult for the same model to be repurposed. Supervised learning encourages the model to learn features heavily correlated with certain labels rather than learn general features of the data itself. This results in highly specialized models that only perform well on the tasks for which they were designed. SSL, on the other hand, is significantly better at generalizing to a variety of tasks [10].

Currently in the medical field, the most common machine learning paradigms used are SSL and transfer learning [5]. For example in a previous paper, a siamese convolutional neural network (CNN) was used to learn embeddings of MRI scans where pairs of images from the same patient were pushed together, and pairs of images from different patients were pushed apart [12]. In another study, self-supervised and semi-supervised learning were combined to pre-train a network to segment brain tumours from MRI scans [1].

2.1.7 Evaluation metrics

One drawback to SSL is that it can be difficult to assess the performance of the model. A common approach is to use the representations learned from the pretext task as inputs to a downstream task and measure its performance. There are two methods for doing so: fine-tuning and linear evaluation [9].

An imaging model can be thought of as a composition of two parts: an encoder and a classifier. In end-to-end fine-tuning, all the weights of both the encoder and classifier are unfrozen and are used as an initialization for training a new model, where all the weights will be updated through supervised learning of the downstream task. In linear evaluation, a small linear classifier is trained on top of the pretext network to perform the downstream task, while the weights of the rest of the network are left frozen [16][10]. In the SimCLR paper for example, the linear evaluation protocol is employed and test accuracy is used as a measure of representation quality [4].

For the specific downstream task of medical image segmentation, recent studies have shown that there is statistical bias present in its evaluation. There has been a trend of cherry-picking improper metrics to fabricate high accuracy scores. This has led to clinical research teams doubting the usefulness of image segmentation models in an clinical setting. One cause for this is the large class imbalance in medical images. The region of interest usually takes up a small fraction of the overall image. For example, in tumour segmentation from an MRI brain image, the majority of the image typically corresponds to the negative class given that the tumour occupies a small region relative to the entire brain. It is also important to use multi-class metrics rather than binary metrics to prevent highly biased results. Due to their direct impact on subsequent diagnoses and treatment, the evaluation of the correctness of medical image segmentation models is critical. [17] proposes the following guidelines for evaluating medical image segmentation models. First, use

the dice similarity coefficient as the primary performance metric. Additionally, provide the intersection-over-union, sensitivity, and specificity. For multi-class problems, these metrics must be calculated for each class individually. It is also important to compare sample visualizations of labelled and predicted segmentations for visual evaluation.

Note that instead of assessing the model based on its downstream performance, it is also possible to evaluate the quality and diversity of the representations using techniques such as nearest-neighbours or clustering [9].

2.2 Interpretability

2.2.1 What is explainable AI?

Most machine learning and deep learning models are considered "black boxes" due to the complex, non-linear, and uninterpretable underlying structures, and many AI systems are unable to explain their decisions to human users [28][7].

Explainable AI (XAI) involves developing techniques to understand the reasoning behind the decision-making process of a machine learning model [23]. An XAI system should be able to explain its understanding, what it has done, what it is currently doing, and what will happen next [7]. The goal of XAI is to create systems that are interpretable to humans while maintaining the degree of accuracy offered by deep learning models. Two questions an explainability method should answer are how and why the model produces its predictions or inferences [28].

There is currently no universal means to measure whether an XAI system is more intelligible to a human user, nor is there a consensus on the definition of an explanation or how it should be assessed [7][28]. Researchers are striving to develop universal, objective criteria for how to assess a model's explanations [28]. [27] considers an explanation to be good if it provides insight on how a model came to its decisions and makes said decisions human-understandable.

The exact definitions of the explainability, understandability, and interpretability of systems are still in debate [20]. In fact, their definitions are dependent on the domain. For example, the background knowledge and needs of the end users will vary depending on their domain [7]. Explainability and interpretability are often used interchangeably in the AI community. In this review we will use the terms "interpretability" and "explainability" interchangeably and define both as the ability to provide a human-interpretable explanation for the decisions made by an AI system.

The most common form of XAI in medical image analysis is saliency mapping, which is a form of visual explanation. Saliency maps depict the most important parts of an image for a decision (such as classification). Most saliency mapping technique use a back-propagation-based approach, while some user a perturbation-based or multiple instance learning-based approach [27].

2.2.2 Interpretability techniques

Existing XAI methods designed for models that use a single image as input are not well-suited for dual-input contrastive learning and fail to account for the interdependent and interactive inputs of contrastive models [23]. While significant progress has been made in the field of XAI in explaining the inner workings of image classification models, limited progress has been made in explaining contrastive learning models.

The following describes a few popular techniques in XAI.

SmoothGrad

SmoothGrad is a simple technique that improves the interpretability of gradient-based saliency maps [26]. It works by generating multiple versions of the same image by adding noise to each one, and then averages the resulting saliency maps of each image. This produces a single saliency map that highlights the importance of each pixel in the model's prediction.

Occlusion Sensitivity

Occlusion sensitivty (also called conditional occlusion in [23]) is a technique used for models such as CNNs that attempts to determine whether the network is actually using the object in the image to make its predictions, or just the surrounding context [31]. It systematically blocks out regions of the input image and then compares the model output. If the model is actually using the object in the image, the class probability should drop when the object is occluded.

Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a technique used for CNNs to visualize important regions of an input image [24]. It uses global average pooling on the gradients flowing into the final convolutional layer to produce a heatmap showing the model's areas of focus.

The following outlines techniques proposed for contrastive learning in [23].

Averaged Transforms

Averaged transforms is an extension of SmoothGrad that attempts to determine the impact that various strengths of data augmentations have on the similarity between two images. Each augmentation applied to an image in contrastive learning has a random level of strength (for example, the amount of rotation). Each transformation T is decomposed into $[t_1, \ldots, t_Z]$ where t_z is a specific strength of T. Let I_1 and I_2 be two input images and let $t_z(I_2)$ be the zth transformation of I_2. The cosine similarity is calculated between I_1

and $t_z(I_2)$ for all z. The gradient of the similarity score with respect to I_1 and I_2 is then computed. Finally, the gradients are averaged across different strength for each image to produce a saliency map.

Pairwise Occlusion

Pairwise occlusion is an extension of occlusion sensitivity that considers simultaneous perturbations of two images. The idea is that a region of an image is considered significant if the model's prediction cannot be deduced from the context surrounding that region. One potential issue with this method is that testing every combination of regions between two images would result in a significantly large increase in the number of necessary computations. This number is reduced by randomly choosing a different size, location, and aspect ratio for the occlusion on each image. Another possible issue is that regions with less significance may be given too much importance. This is alleviated by performing a softmax operation on the scores.

Interaction-CAM

Interaction-CAM is an extension of Grad-CAM that considers the joint activations and gradients of both input images to create explainability maps that highlight their common features. Features are considered significant if they are jointly active in both images.

Chapter 3

Datasets

Three medical imaging datasets from different anatomical regions were used in this study which utilize a variety of imaging techniques, such as MRI, X-ray, and colonoscopy. These datasets were chosen for their diversity, accessibility, and frequent use in research papers. For every dataset, to standardize the input data, the pixel values of each image were normalized.

3.1 BraTS Dataset

The first dataset used for this research comes from the Brain Tumour Segmentation (BraTS) dataset from the University of Pennsylvania's Multimodal Brain Tumour Segmentation Challenge 2020 [18]. This dataset contains 369 clinically-acquired multimodal MRI brain scans featuring glioblastoma and lower grade glioma, along with ground truth labels from board-certified neuroradiologists.

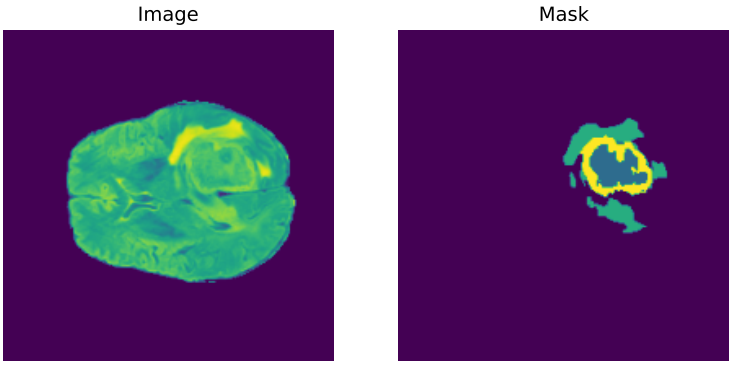


Figure 3.1: Sample image and segmentation mask from the BraTS dataset

Each MRI scan is a 3D image composed of 2D image slices. However, for the purposes of this research and to reduce computational complexity, the median 2D slice was extracted from each volume and used for training.

3.2 Lung Mask Image Dataset

The second dataset is a publicly-available collection of chest X-ray scans from Kaggle [19]. This dataset includes 19051 X-ray scans in total. Each 2D image is paired with a corresponding binary segmentation mask, marking the shape of the lungs in the image.

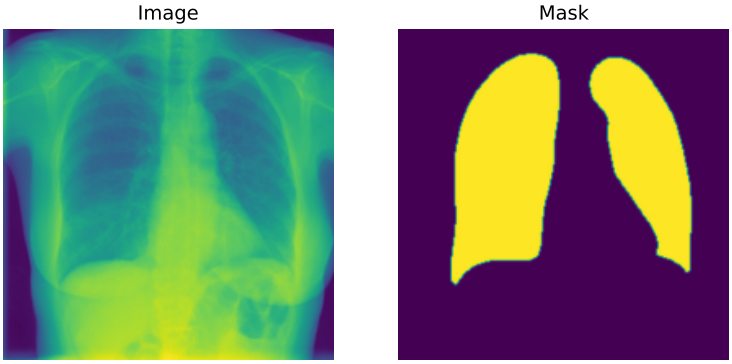


Figure 3.2: Sample image and segmentation mask from the Lung Mask Image dataset

3.3 Kvasir-SEG Dataset

The third dataset comes from the Simula Research Laboratory and is composed of 1000 gastrointestinal images with corresponding segmentation masks marking polyps [13]. These images were taken using a colonoscope and the segmentation masks were drawn by an experienced gastroenterologist.

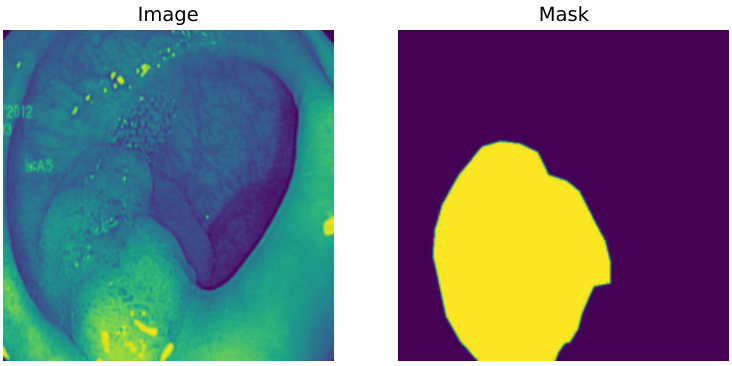


Figure 3.3: Sample image and segmentation mask from the Kvasir-SEG dataset

Chapter 4

Image Segmentation Task

For the first part of this study, medical image segmentation was performed using self-supervised learning. This self-supervised model would be the foundation for subsequent research. For the research performed in this chapter, the BraTS dataset was used.

4.1 Methods

4.1.1 Supervised Model (Baseline)

The first objective in this study was to perform image segmentation using a supervised model, which would later serve as a comparative baseline against the self-supervised model.

For the supervised learning task, a UNet model was implemented. The encoder is comprised of four sequential blocks, each consisting of a double convolutional layer with ReLU activation functions. Between each encoder block, a max pooling layer is applied. The decoder consists of three upsampling blocks. The final layer maps the output of the decoder to four output channels, each representing a class in the segmentation mask.

The UNet model was trained on 329 samples (with 40 validation samples) for 100 epochs with a batch size of 64. The ADAM optimizer was used with a polynomial decay learning rate schedule in which the learning rate decreased according to $\left(1 - \frac{\text{epoch}}{\text{total\_epochs}}\right)^{0.9}$ with an initial learning rate of 0.001. A custom dice loss function was used.

4.1.2 Self-Supervised Model

To implement contrastive learning, a pretrained MoCo model was adapted. The MoCo model was chosen due to its superior computational efficiency in comparison to frameworks such as SimCLR, and the availability of pretrained models [8][22]. As a preprocessing step, a random transformation, (rotation, horizontal flip, or

vertical flip), was applied to each sample in the dataset to augment the data and create the sets of positive pairs.

The MoCo model uses a ResNet-50 base encoder. Since the MRI images are grayscale, the first convolutional layer was adapted for single-channel inputs. The MoCo architecture is comprised of two encoders, the key encoder and query encoder, along with an MLP head. The pretrained MoCo model was fine-tuned on the augmented BraTS dataset using a stochastic gradient descent optimizer with learning rate 0.03 and momentum 0.9. It was trained for 100 epochs with a batch size of 4.

For the downstream task, the same UNet model was used as the supervised approach but the encoder was replaced with the ResNet-50 backbone from the query encoder of the MoCo model. The same optimizer and scheduled learning rate were used as the supervised model.

4.2 Results

The performance of both models was evaluated based on four metrics: Dice score, IoU, sensitivity, and specificity. These metrics were calculated for each class and then averaged.

	Dice score	**IoU**	**Sensitivity**	**Specificity**
Training	0.7250	0.6823	0.8423	0.9707
Validation	0.7370	0.6968	0.8376	0.9739

Table 4.1: Supervised model results on the training and validation sets

	Dice score	**IoU**	**Sensitivity**	**Specificity**
Training	0.8886	0.8760	0.9638	0.9835
Validation	0.8804	0.8729	0.9759	0.9787

Table 4.2: Self-supervised model results on the training and validation sets

The self-supervised MoCo model outperformed the supervised UNet model in all metrics across both the training and validation sets. This indicates that the features learned from the MoCo model were highly beneficial in segmenting the brain tumours. Both models had high validation scores with respect to the training scores, indicating that they were able to generalize well to unseen data.

The results of this task provide compelling evidence of the efficacy of self-supervised learning in medical imaging applications, motivating the subsequent research in this study.

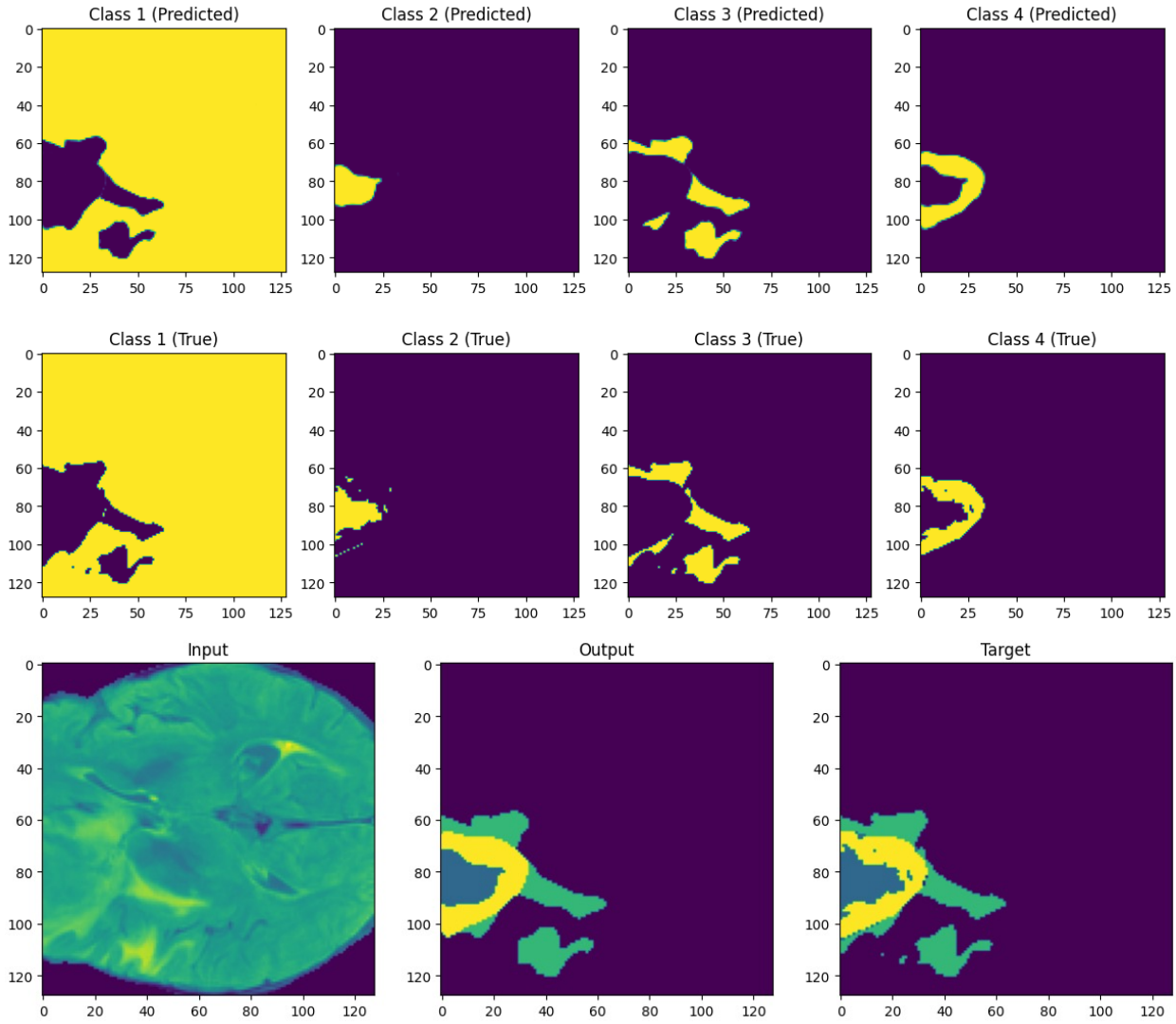


Figure 4.1: Sample result from self-supervised model. Top row: Model's prediction per class. Middle row: True segmentation mask per class. Bottom row (left to right): Input image, model's output, true mask.

Chapter 5

Saliency Maps

5.1 Selecting an Interpretability Technique

To better understand the decisions made by the self-supervised model, eight interpretability techniques for contrastive learning were applied to positive pairs of input images using the fine-tuned MoCo model. These techniques were adapted from a publicly available GitHub repository[1] provided by [23]. See section 2.2.2 for descriptions of each technique.

Fig. 5.1 shows the eight interpretability techniques for a sample input image from the BraTS dataset and its augmented version. For the remainder of this research, we will focus on rotation as the augmentation strategy.

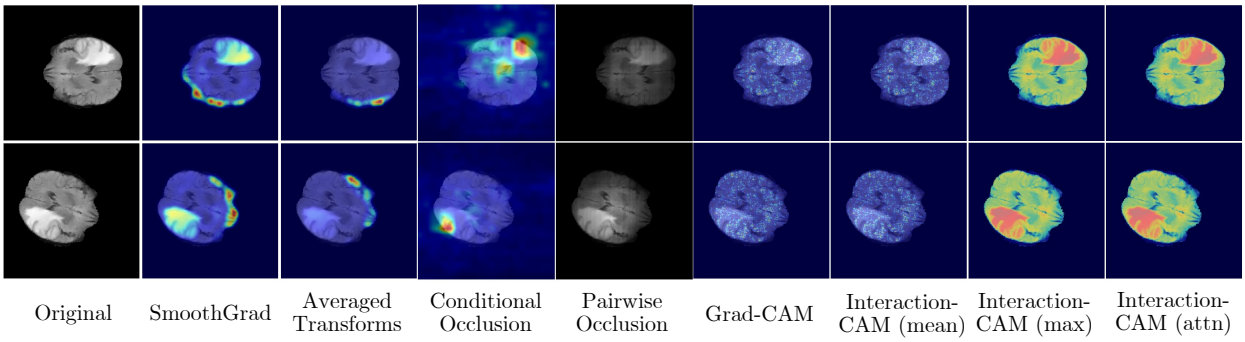


Figure 5.1: Various interpretability techniques applied to a sample input from BraTS dataset. Top row: Original image. Bottom row: Rotated image. From left to right: Original image, SmoothGrad, Averaged Transforms, Conditional Occlusion, Pairwise Occlusion, Grad-CAM, variations of Interaction-CAM.

Averaged transforms, pairwise occlusion, Grad-CAM, and Interaction-CAM (mean) did not provide any meaningful information about the importance of various regions of the input image in the model's prediction because they did not highlight any regions at all. Interestingly, SmoothGrad highlighted an area very

[1] https://github.com/fawazsammani/explain-cl

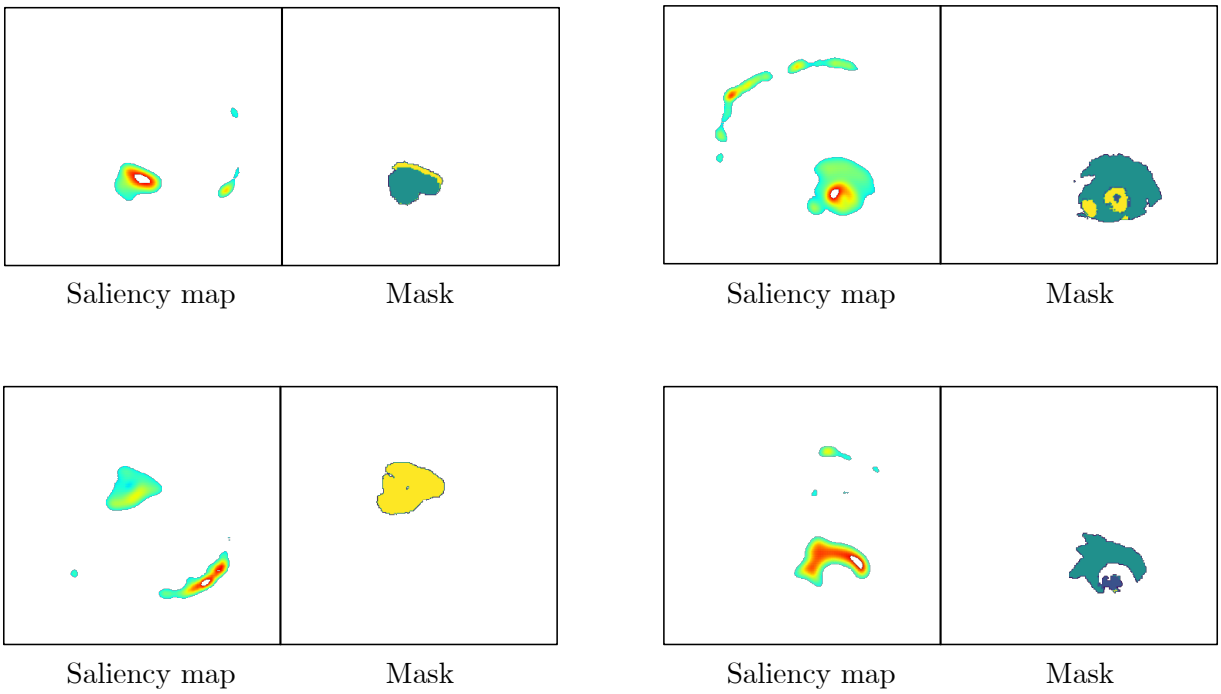


Figure 5.2: Saliency maps generated using SmoothGrad with corresponding segmentation masks for four samples of the BraTS dataset.

similar to the shape of the true segmentation mask. To investigate this further, saliency maps using SmoothGrad were generated for numerous input images from the BraTS dataset and compared visually with the segmentation masks, see Fig. 5.2.

It is hypothesized that the saliency maps, which represent the regions of the input image that the model deemed significant when generating its output, are more meaningful when they correspond with the ground truth segmentation masks. This is because the area highlighted by the segmentation mask is the defining region of interest of the image, or in other words the feature that makes an image distinct. For example, in the case of the BraTS dataset, the main feature of an image is the brain tumor outlined by the segmentation mask. If the saliency map highlights this tumor area, it implies that this interpretability technique is accurately depicting the area that the model considered significant since the model performed well on the segmentation task.

Hypothesis 1 (H1): *The model produces more interpretable results when the saliency maps, which represent the regions of the input image that were important for the model's prediction, are more similar to the ground truth segmentation masks (which are the relevant regions of the image).*

For the remainder of this research, we will move forward with the SmoothGrad method as the interpretability technique because it most accurately highlights the region corresponding to the segmentation mask.

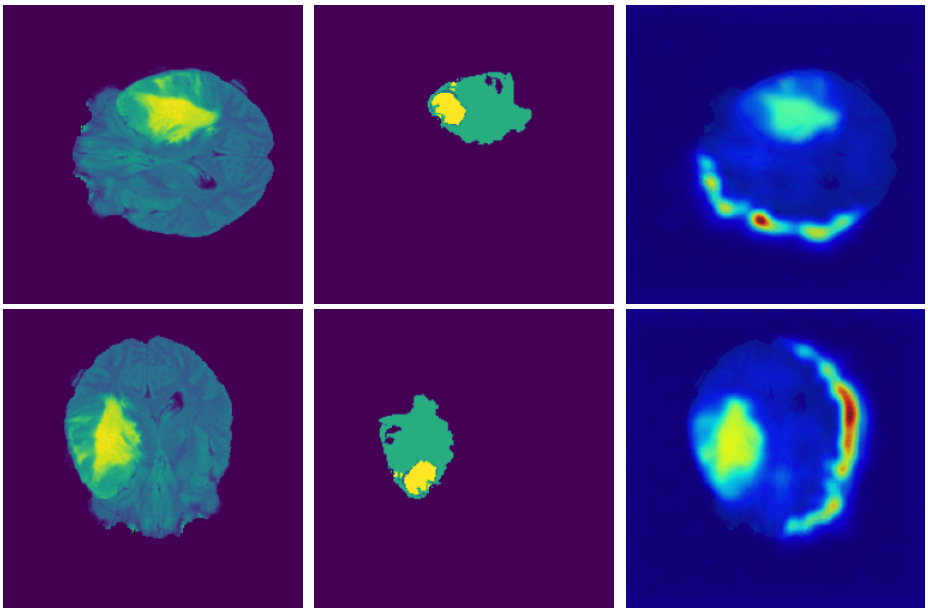


Figure 5.3: Comparison of original images from BraTS dataset, segmentation masks, and saliency maps. Top row: Original orientation. Bottom row: Rotated by 95°.

5.2 Measuring Interpretability

A question that arises from H1 is: since the contrastive learning model uses augmentation to create positive pairs, are there certain amounts of augmentation that optimize this measure of interpretability?

The next objective is to measure the model's interpretability as a function of the augmentation applied to the image. Using H1, we will proceed to examine whether the interpretability of the model is constant with the amount of augmentation (rotation) applied to the input image, or if an interesting pattern emerges.

5.3 Generating Saliency Maps

To accomplish this, 75 images from the BraTS dataset were selected. For each degree θ from 0 to 360, each sample in the batch was rotated by θ and a SmoothGrad saliency map was generated for the rotated image, resulting in a total of 75×360 saliency maps produced. Each saliency map was then rotated by $-\theta$ and compared to the corresponding ground truth segmentation mask using the Dice score in Eq. 5.1. For each θ, the Dice scores across the batch were averaged, and then plotted against the rotation angle.

$$\text{Dice} = 2 \times \frac{\text{intersection}}{\text{intersection} + \text{union} + \epsilon} \tag{5.1}$$

The pre-trained MoCo model was also fine-tuned o the Lung Mask Image and the Kvasir-SEG datasets separately, with the plot generation procedure repeated for each. All three plots can be found in Table 6.1.

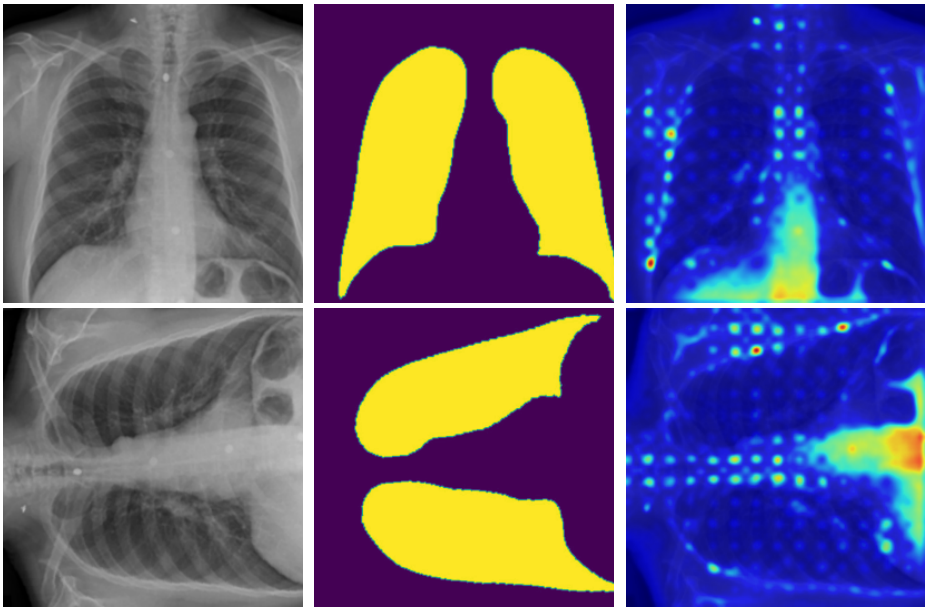


Figure 5.4: Comparison of original images from Lung Mask Image dataset, segmentation masks, and saliency maps. Top row: Original orientation. Bottom row: Rotated by 95°.

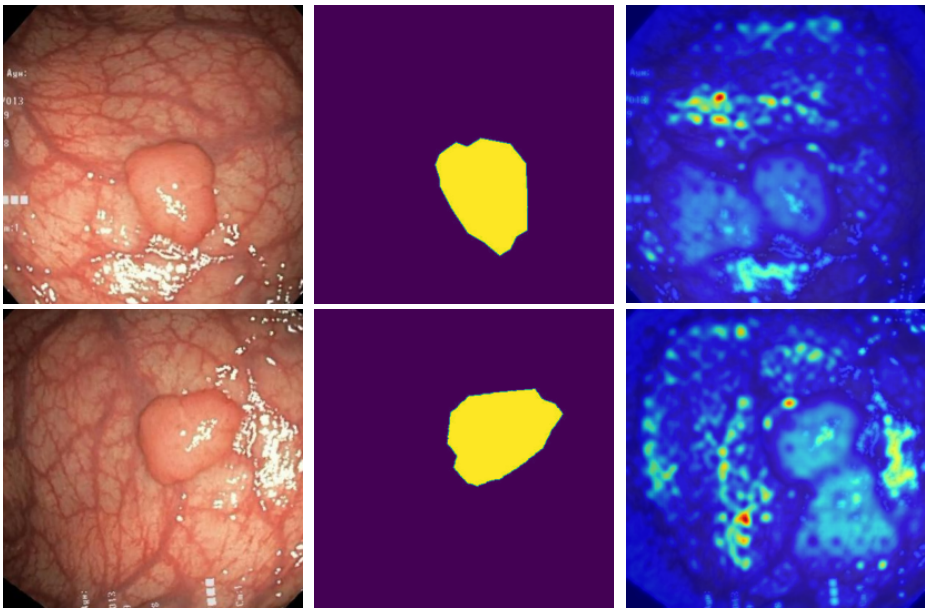


Figure 5.5: Comparison of original images from Kvasir-SEG dataset, segmentation masks, and saliency maps. Top row: Original orientation. Bottom row: Rotated by 95°.

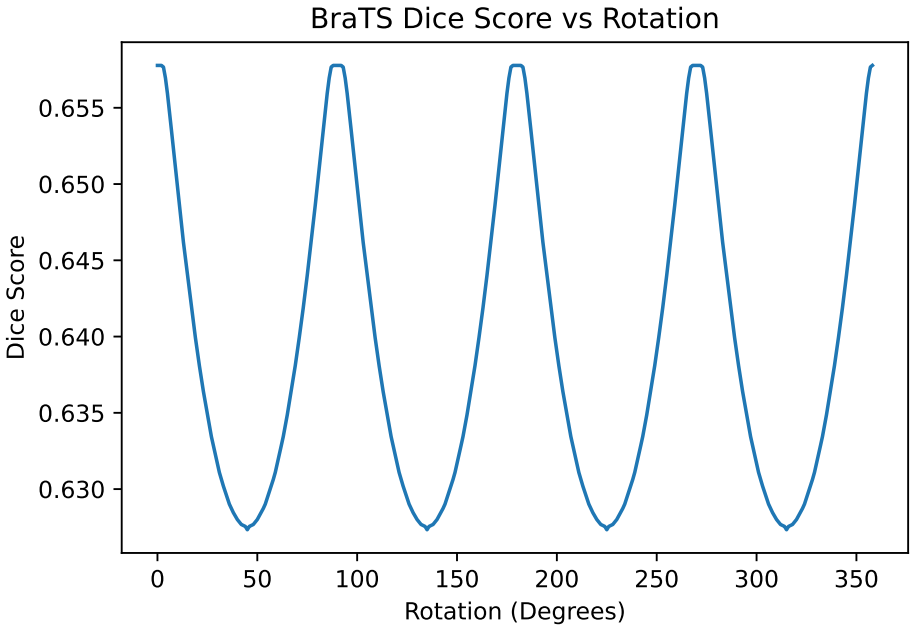


Figure 5.6: Average Dice score vs. rotation angle for BraTS dataset

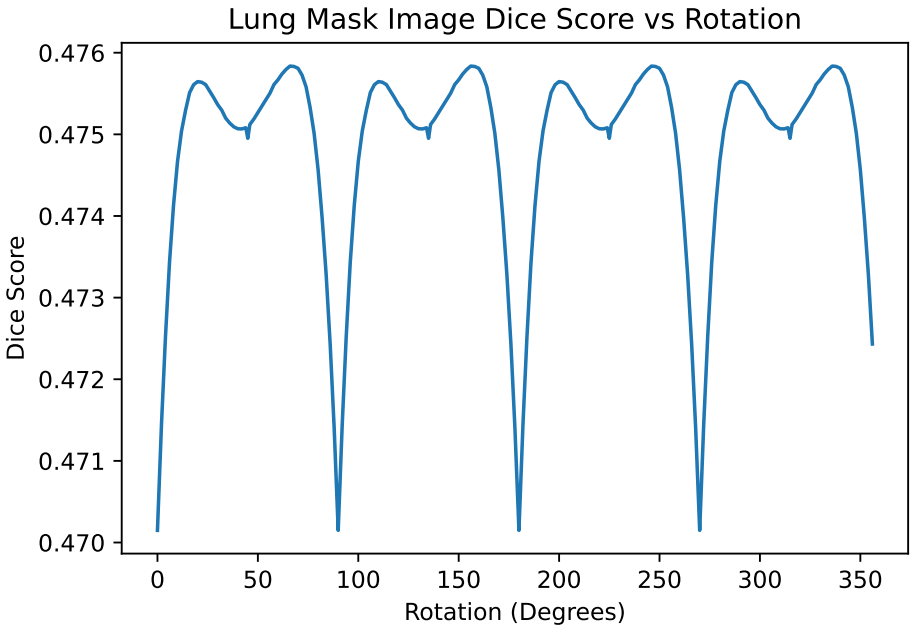


Figure 5.7: Average Dice score vs. rotation angle for Lung Mask Image dataset

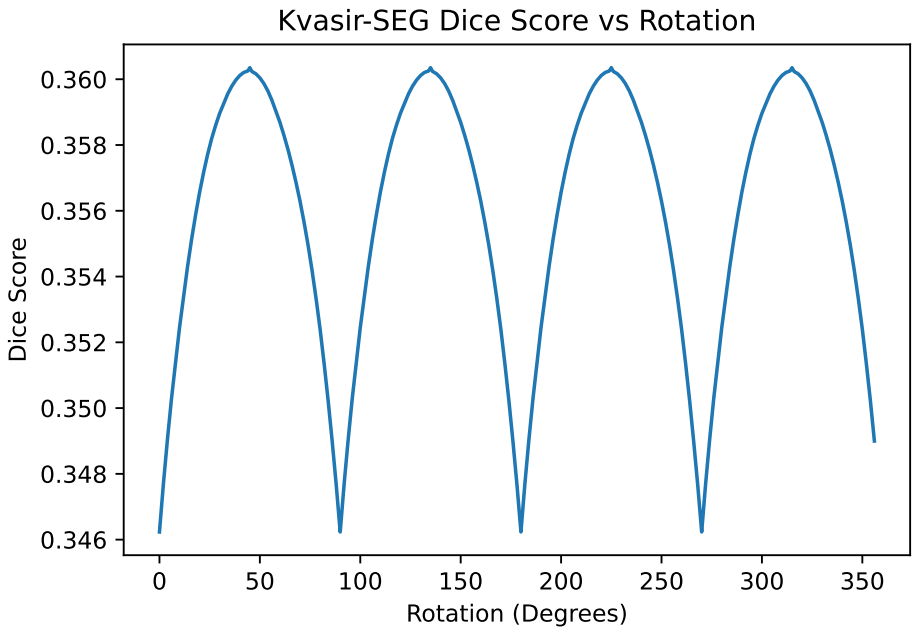


Figure 5.8: Average Dice score vs. rotation angle for Kvasir-SEG dataset

In all three plots, an interesting pattern emerges when comparing the effectiveness of the saliency maps versus the amount of rotation in the input image.

The Dice score versus rotation angle for the BraTS (brain imaging) dataset in Fig. 5.6 exhibits periodic behaviour. Dice scores oscillated between approximately 0.627 and 0.658 with peaks around every 90° of rotation. The maxima occur at 90°, 180°, and 270°. The minima occur at 45°, 135°, 225°, and 315°.

The plot for the Lung Mask Image dataset in Fig. 5.7 reveals another periodic shape with a different amplitude range, namely between approximately 0.470 and 0.476. For this dataset the maxima occur at 66°, 156°, 246°, and 336° degrees. The minima occur at 90°, 180°, and 270°.

Another periodic shape is presented in the plot for the Kvasir-SEG (GI tract) dataset in Fig. 5.8, this time oscillating between Dice scores of approximately 0.346 and 0.360. The maxima of this plot occur at 45°, 135°, 225°, and 315°. Similarly to the chest dataset, the minima occur at 90°, 180°, and 270°.

Chapter 6

SVM Classifier

In Figs. 5.6, 5.7, and 5.8, extrema occur at 90°, 180°, and 270°. For the BraTS dataset, these values correspond to the local maxima, and for the other two datasets, these values correspond to the local minima. According to our hypothesis, this means that the model is either most or least interpretable when the input image is rotated by a multiple of right angles.

One explanation for this could be due to the natural symmetric or orientation-specific features of the input images. For example, the images from the BraTS dataset (see Fig. 3.1) are approximately symmetric about the x-axis. When the images are rotated at 90° intervals, the symmetric and orientation-specific features may be more easily recognizable. During fine-tuning, the MoCo model may have learned representations that capture these features, which could result in more understandable visualizations.

6.1 Plotting HOG Features

We hypothesize that the quality of representations learned by the model deteriorates when the network takes "shortcuts" using orientation-specific features of the images to match positive pairs. For instance, the model may rely on Histogram of Oriented Gradients (HOG)-like features, as evidenced by work dating back to [14] and [6]. These studies have shown that HOGs effectively characterize the appearance of objects, which may lead the model to depend on such features.

Hypothesis 2 (H2): *The representations learned by the model are worse when it relies on orientation-specific features.*

Examples of features extracted using the HOG method for each dataset are plotted in Fig. 6.1.

6.2 SVM Classifier

To test whether the MoCo model is using the HOG-like features as "shortcuts", a support vector machine (SVM) was used to investigate how changes in orientation via rotation impact the model's ability to recognize different orientations of the same image. The SVM was trained to classify the HOG features of images rotated by θ vs. unrotated. The parameters for the SVM were selected using cross-validation. The accuracy of the SVM was then plotted against the rotation angle θ. If the SVM shows varying accuracy at the same angles as the MoCo model, it could indicate that the MoCo model is using orientation-specific features when learning its representations.

The results are in Table 6.1. For all three datasets, the SVM classification accuracy is low for θ close to $0°$ and (and $360°$) and high everywhere else, reflecting the increased difficulty of determining whether an image was rotated when it has only been augmented slightly.

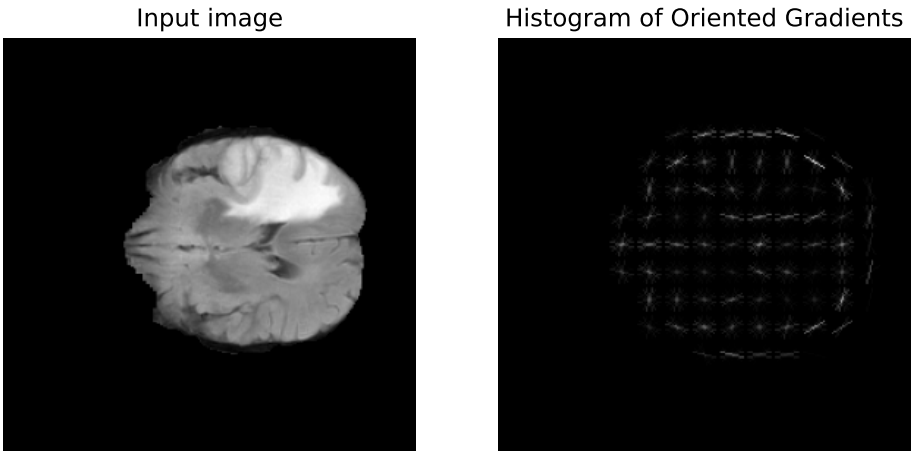
For the BraTS dataset, the accuracy remains very high from $\theta = 11$ to $\theta = 348$, with a single outlier at $\theta = 180$. For the Lung Mask Image dataset, the accuracy is high from $\theta = 15$ to $\theta = 341$, with slight local minimia at $\theta = 90$, $\theta = 180$, and $\theta = 270$. Similarly, the Kvasir-SEG dataset has a high accuracy from $\theta = 12$ to $\theta = 340$, with slight local minima at $\theta = 90$, $\theta = 180$, and $\theta = 270$.

This means that the SVM had a bit of trouble determining whether an image was rotated or not based on its HOG at $90°$ rotation intervals, with the most difficulty at about $343°$ to $13°$.

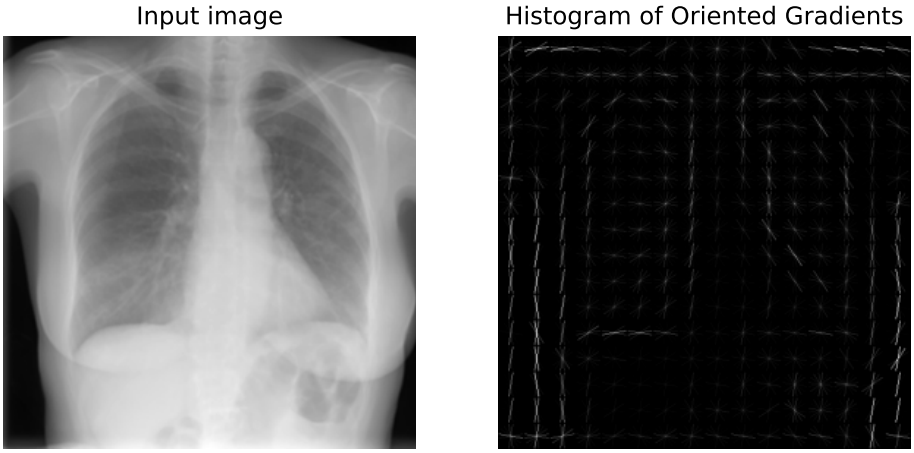
6.3 Self-Supervised Model vs. SVM

For the Lung Mask Image and Kvasir-SEG datasets, the SVM accuracy plots somewhat align with the Dice score plots in Table 6.1 in that their local minima occur at the same rotation angles. However, unlike the Dice score plots, the accuracy plots do not exhibit a regular periodic pattern, suggesting that the classifier's performance is generally rotation invariant except for at a few specific angles.

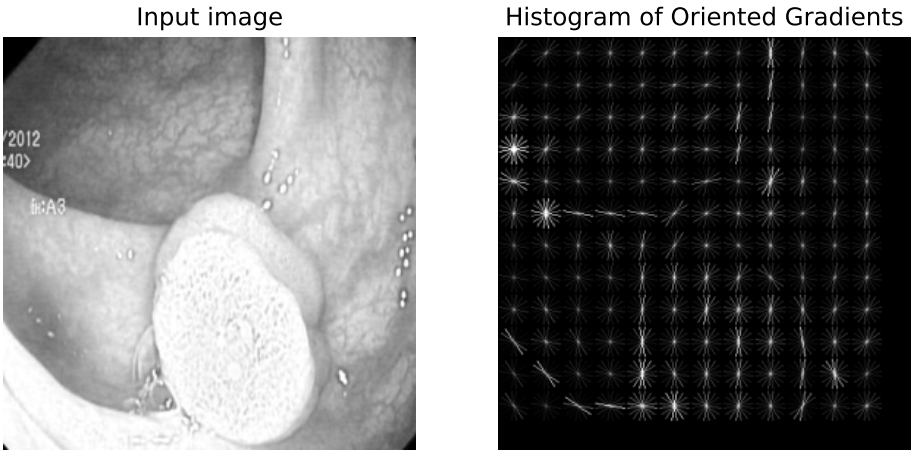
The outlier at $\theta = 180$ in the BraTS SVM accuracy plot does not correspond with a local minima in the BraTS dice score plot. Furthermore, the interpretability of the MoCo model trained on the BraTS data is at a maximum when the accuracy of the SVM classifier is at a minimum.



(a) BraTS dataset



(b) Lung Mask Imaging dataset



(c) Kvasir-SEG dataset

Figure 6.1: Sample HOG features for each dataset.

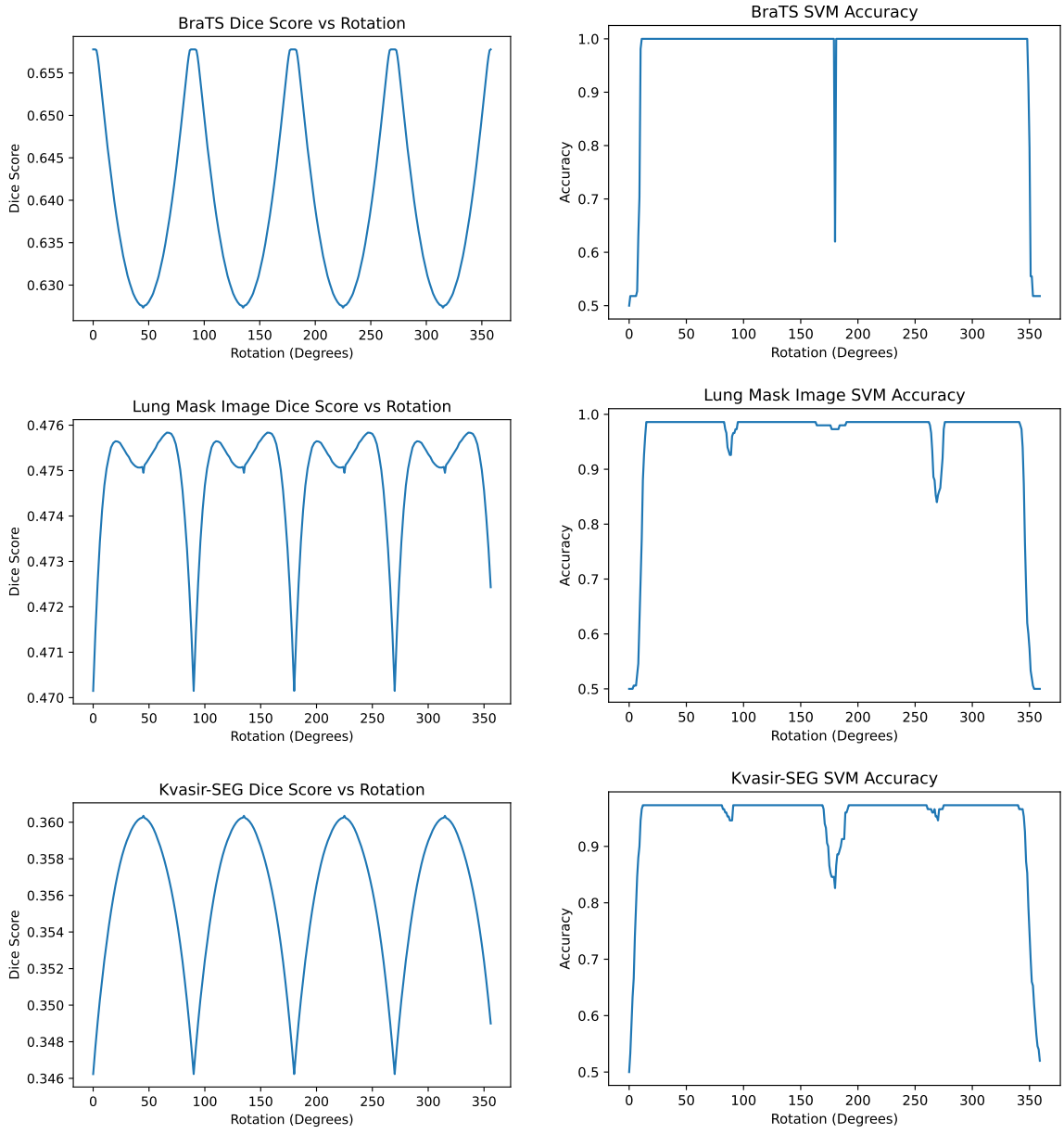


Table 6.1: Average correspondence (using Dice score) between ground truth segmentation and saliency map obtained using trained features vs. rotation angle (left) and accuracy of an SVM classifier classifying original images versus rotated image vs. rotation angle (right).

Chapter 7

Discussion

In Chapter 4, we trained a supervised UNet model and fine-tuned a self-supervised pre-trained MoCo model on the BraTS dataset. It was found that the MoCo model outperformed the UNet model, demonstrating the effectiveness of self-supervised learning for medical imaging.

In Chapter 5, we saw that SmoothGrad saliency map had the highest accuracy in highlighting the region captured in the ground truth segmentation mask. We proposed that saliency maps that are more similar to the ground truth correspond with a more accurate interpretability technique. We plotted the interpretability of the model, measured by the Dice score, against a range of increasing image rotations from 0 to 360 degrees and noticed a periodic function with a period of 90 for all three datasets.

In Chapter 6, we hypothesized that the MoCo model may be using HOG-like features when learning its representations, resulting in higher interpretability when the input image was rotated by a 90° interval. However, when we trained an SVM to classify HOG features of an image as rotated vs. unrotated, the plots did not directly align with the Dice score plots.

The key claims to be drawn from this research include the following:

1. Using the learned representations from a fine-tuned self-supervised model is highly effective in improving the accuracy of medical image segmentation of the BraTS dataset.

2. When the similarity between the saliency maps and segmentation masks is plotted against increasing angles of rotation θ, numerous datasets exhibit periodic behaviour with a period of 90°.

3. Some datasets have local minima at intervals of $\theta = 90$, whereas others have local maxima at those points.

4. The values of θ with the lowest interpretability do not necessarily correspond with values of θ where an SVM would struggle to classify the image as rotated vs. unrotated, implying that the MoCo model does not rely on HOG-like features to create its representations.

These claims are significant in the context of the interpretability of contrastive learning models used for medical imaging because they show that the interpretability of the model exhibits a distinct periodic pattern as the augmentation increases. This means that for certain angles of rotation, the interpretability of the model is "better" in the sense that the interpretability technique aligns more closely with the ground truth segmentation. This is a very interesting result because it provides evidence that there are optimal augmentations at repeating intervals when training a self-supervised model for interpretability and that the model is not equally interpretable with any amount of augmentation. These findings offer valuable insight into selecting the optimal amount of augmentation to maximize interpretability. This work also shows that these optimal augmentations are dependent on the dataset, but the interval at which the optimal augmentations repeat is the same.

Chapter 8

Conclusion and Future Work

In this research, an interesting pattern was observed after fine-tuning a self-supervised MoCo model on medical imaging datasets. As the angle of rotation of the input image was increased, the interpretability of the model—measured by the similarity between SmoothGrad saliency maps and ground truth segmentation masks—exhibited periodic behaviour, with local maxima at intervals of 90°.

Moving forward, there is significant potential for future research to build off of the findings of this work. First, it's necessary to test several other datasets before we can claim that the periodic patterns we observed in our experiments are common. It would be interesting to explore both medical and natural images to see if this pattern exists across various domains. Next, our experiments used rotation angle as the independent variable, but it's important to test our findings using other augmentations that allow for continuous variation. Some examples of such augmentations include Gaussian blur, cropping, and brightness scaling. If interesting patterns arise for many augmentations, it could provide significant insight into augmentation choice when training contrastive learning models for interpretable results.

Bibliography

[1] Varghese Alex et al. "Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation". In: *J. Med. Imaging* (2017). DOI: https://doi.org/10.1117/1.JMI.4.4.041311.

[2] Yamen Ali et al. "Self-Supervised Learning for 3D Medical Image Analysis using 3D SimCLR and Monte Carlo Dropout". In: *Proceedings of Machine Learning Research* 1.6 (2021). DOI: https://doi.org/10.48550/arXiv.2109.14288.

[3] Liang Chen et al. "Self-supervised learning for medical image analysis using image context restoration". In: *Medical Image Analysis* 58 (2019). DOI: https://doi.org/10.1016/j.media.2019.101539.

[4] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning* (2020). DOI: https://doi.org/10.48550/arXiv.2002.0570.

[5] Alexander Chowdhury et al. "Applying Self-Supervised Learning to Medicine: Review of the State of the Art and Medical Implementations". In: *Informatics* (2021). DOI: https://doi.org/10.3390/informatics8030059.

[6] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee. 2005, pp. 886–893.

[7] David Gunning et al. "XAI—Explainable artificial intelligence". In: *Science Robotics* 4.37 (2019). DOI: http://dx.doi.org/10.1126/scirobotics.aay7120.

[8] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: (2020). DOI: https://doi.org/10.48550/arXiv.1911.05722.

[9] *How do you design effective pretext tasks for self-supervised learning of neural networks?* URL: https://www.linkedin.com/advice/3/how-do-you-design-effective-pretext.

[10] Shih-Cheng Huang et al. "Self-supervised learning for medical image classification: a systematic review and implementation guidelines". In: *npj Digital Medicine* 6.74 (2023). DOI: https://doi.org/10.1038/s41746-023-00811-0.

[11] Ashish Jaiswal et al. "A Survey on Contrastive Self-Supervised Learning". In: *Technologies* (2021). DOI: https://doi.org/10.3390/technologies9010002.

[12] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. "Self-supervised Learning for Spinal MRIs". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (2017), pp. 294–302. DOI: https://doi.org/10.48550/arXiv.1708.00367.

[13] Debesh Jha et al. "Kvasir-seg: A segmented polyp dataset". In: *International Conference on Multimedia Modeling*. Springer. 2020, pp. 451–462.

[14] David G Lowe. "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[15] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: (2013). DOI: https://doi.org/10.48550/arXiv.1301.3781.

[16] Ishan Misra. *Self-Supervised Learning - Pretext Tasks*. URL: https://ebetica.github.io/pytorch-Deep-Learning/en/week10/10-1/.

[17] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. "Towards a guideline for evaluation metrics in medical image segmentation". In: *BMC Research Notes* 15.210 (2022). DOI: https://doi.org/10.1186/s13104-022-06096-y.

[18] *Multimodal Brain Tumor Segmentation Challenge 2020: Data*. URL: https://www.med.upenn.edu/cbica/brats2020/data.html.

[19] Neelkant Newra. *Lung Mask Image Dataset*. 2022. DOI: 10.34740/KAGGLE/DSV/3559777. URL: https://www.kaggle.com/dsv/3559777.

[20] Simon J.D. Prince. *Understanding Deep Learning*. The MIT Press, 2023. ISBN: 9780262048644.

[21] Veenu Rani et al. "Self-supervised Learning: A Succinct Review". In: *Archives of Computational Methods in Engineering* 30 (2023), pp. 2761–2775. DOI: https://doi.org/10.1007/s11831-023-09884-2.

[22] Meta Research. *MoCo: Momentum Contrast for Unsupervised Visual Representation Learning*. URL: https://github.com/facebookresearch/moco.

[23] Fawaz Sammani, Boris Joukovsky, and Nikos Deligiannis. "Visualizing and Understanding Contrastive Learning". In: *IEEE Transactions on Image Processing* (2023). DOI: https://doi.org/10.48550/arXiv.2206.09753.

[24] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[25] Deval Shah and Abhishek Jha. *Self-Supervised Learning and Its Applications*. URL: https://neptune.ai/blog/self-supervised-learning.

[26] Daniel Smilkov et al. “SmoothGrad: removing noise by adding noise”. In: (2017). DOI: https://doi.org/10.48550/arXiv.1706.03825.

[27] Bas H.M. van der Velden et al. “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis”. In: *Medical Image Analysis* 79 (2022). DOI: https://doi.org/10.1016/j.media.2022.102470.

[28] Giulia Vilone and Luca Longo. “Notions of explainability and evaluation approaches for explainable artificial intelligence”. In: *Information Fusion* 76 (2021), pp. 89–106. DOI: https://doi.org/10.1016/j.inffus.2021.05.009.

[29] Wei-Chien Wang et al. “A Review of Predictive and Contrastive Self-supervised Learning for Medical Images”. In: *Machine Intelligence Research* (2023). DOI: https://doi.org/10.48550/arXiv.2302.05043.

[30] Jiashu Xu. “A Review of Self-supervised Learning Methods in the Field of Medical Image Analysis”. In: *International Journal of Image, Graphics and Signal Processing* 4 (2021), pp. 33–46. DOI: http://dx.doi.org/10.5815/ijigsp.2021.04.03.

[31] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *Computer Vision–ECCV 2014*. Ed. by David Fleet et al. Vol. 8689. Lecture Notes in Computer Science. Cham: Springer, 2014. DOI: 10.1007/978-3-319-10590-1_53.