**(1) How Vision Transformers (ViTs) and Spatial Transformers can be used for early identification of IBDs (e.g., Ulcerative Colitis, Crohn's Disease)?**

**Vision Transformers (ViTs)** and **Spatial Transformer Networks (STNs)** offer advanced approaches for the early identification of inflammatory bowel diseases (IBDs) such as Ulcerative Colitis and Crohn's Disease by improving the ability to extract key visual features from medical images, such as endoscopic or histopathological images. Here’s how each can be utilized:

* **Vision Transformers (ViTs):** ViTs divide an image into patches and treat each patch as a sequence (like tokens in natural language models), allowing the model to capture relationships between various regions of the image. This ability to focus on the spatial relationships across the entire image makes ViTs effective for identifying subtle patterns associated with IBD, such as mucosal irregularities or inflammatory signs in endoscopic images.

For IBD identification, ViTs could:

* + Analyze endoscopy or colonoscopy images, detecting early signs of inflammation, ulcers, or abnormal tissue.
  + Learn from both local and global image features to identify complex patterns that may indicate disease.
  + Offer robust performance in detecting subtle, early-stage symptoms that might be missed by other methods.
* **Spatial Transformers (STNs):** STNs can be integrated into a neural network architecture to learn spatial transformations like translation, rotation, and scaling directly from the data, allowing the network to focus on the most relevant regions of an image. In IBD diagnostics, this can help:
  + Automatically focus on specific regions in a colonoscopy image where inflammation or ulcers are most visible.
  + Correct variations in the camera's orientation or zoom during medical procedures, making the detection process more reliable.

STNs can enhance the diagnostic process by dynamically adjusting the input image to ensure the network focuses on the most important diagnostic regions, regardless of the image's quality or position.

**(2) How are ViTs and Spatial Transformers different from classic neural network approaches?**

* **Vision Transformers vs. Convolutional Neural Networks (CNNs):**
  + **CNNs** rely on convolutional layers to extract local features from small patches of the image. They are highly effective in tasks like medical image analysis but struggle with capturing global relationships between distant image regions.
  + **ViTs**, on the other hand, do not rely on convolutions but treat images as a sequence of patches. This allows ViTs to better capture long-range dependencies between different parts of an image, making them more powerful for tasks where both local and global context matter, such as identifying complex, spatially dispersed symptoms of IBD.
* **Spatial Transformers vs. Classic CNNs:**
  + Traditional CNNs expect input images to be preprocessed and fixed, meaning they are not inherently able to handle variations in scale, rotation, or translation without explicit augmentation.
  + **STNs** add a layer that can learn to transform input images dynamically, improving robustness to image distortions or variations that commonly occur in medical imaging, such as during endoscopic procedures.

**(3) What are the best Vision Transformers and Spatial Transformer models for detecting IBDs?**

Here are some models that could be particularly useful for detecting IBDs:

* **Vision Transformer Models:**
  + **ViT (Vision Transformer)**: The standard ViT model has shown impressive performance in medical imaging tasks due to its ability to capture complex spatial dependencies. It could be directly applied to colonoscopy or biopsy images.
  + **Swin Transformer**: A hierarchical transformer model that processes images at multiple scales, making it highly suitable for detecting subtle details in large medical images, such as differentiating between healthy and inflamed tissue in IBD.
  + **DINO (Self-supervised Vision Transformer)**: DINO uses self-supervised learning to create robust image representations, which could be particularly useful for training on large, unlabeled medical datasets for early IBD detection.
* **Spatial Transformer Models:**
  + **STNs (Spatial Transformer Networks)**: These can be added to a CNN-based or transformer-based model to enhance performance by enabling the model to focus on the most relevant parts of the image, regardless of their position or scale.
  + **Deformable DETR**: This model combines the benefits of transformers with a deformable attention mechanism, allowing it to focus on specific regions of interest in images. This can be beneficial for highlighting specific areas of inflammation or ulceration in IBD patients.

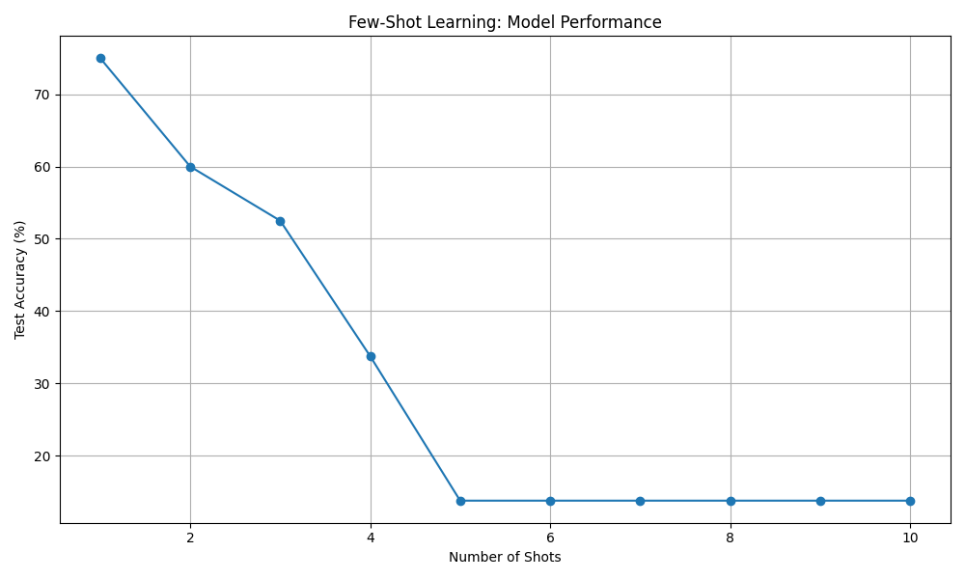
**(3) How can I improve the work of vision transformers using thick data approaches:**

a.) What if I employed Siamese Neural Networks for one-shot or multi-shots learning?

A Siamese neural network (sometimes called a twin neural network) is an artificial neural network that uses the same weights while working in tandem on two different input vectors to compute comparable output vectors. Performing training using One-Shot or Few-Shot learning means that we will utilize a pre-trained Neural network on the Most or some of the categories of the Kvasir Dataset and then utilize a control category which will then be used to perform the One or Few shot learning. In theory, it works the following way: We have a trained network the classifies Dogs and Cats, but then we want it to classify horses as well. The problem is that we have few horse images or even worse we only have a single image of a horse as our dataset. So, we utilize the model to accurately classify what a Dog or a Cat is. And then use the few examples of a Horse to recognize that it is not either of those previous 2 classes: Correctly classifying it as the other(Horse).

To replicate this, we will first train a Siamese Neural Network on all classes on the Kvasir dataset except the normal “polyps” class, after trained we will introduce it in a few shots learning approach and measure model performance at different numbers of shots up to 10.

For the Siamese neural network, we first approached trying to train our own Siamese NN based on the Vision Transformer (ViT) model. This approach had to be scratched as our hardware resources severely limited the possible training/fine tuning of this model to the Kvasir-V2 dataset. We then opted for the ViTMSN model for Image Classification (https://arxiv.org/abs/2204.07141) which is a ViT base masked Siamese neural network. It is pretrained as well, we just had to fine tune it our dataset.

After the model got finetuned, it achieved an accuracy of 83.75% on Testing. We then proceeded to do our single and few shot experiments. Which yielded the following results:

|  |  |
| --- | --- |
| Shots | Accuracy(%) |
| 1 | 75 |
| 2 | 60 |
| 3 | 52.5 |
| 4 | 33.75 |
| 5 | 13.75 |
| 6 | 13.75 |
| 7 | 13.75 |
| 8 | 13.75 |
| 9 | 13.75 |
| 10 | 13.75 |

As we can see, the model performed the best after being trained on a One-shot, which then steadily declined until it reached 5 shots, after which it goes steady at 13.75% accuracy.

The test accuracy starts at 75% with One-shot, then decreases steadily as the number of shots increases. By 4 shots, the accuracy is down to 33.75%, and it remains at 13.75% from 5 shots up to 10 shots.

This indicates that the initial performance with just One-shot is quite high, but the model is not able to effectively leverage the additional few-shot samples to continue improving its accuracy. The accuracy declines significantly as more shots are added.

This is quite different from the trend shown in the original line chart. The results suggest the model may have limitations in generalizing from limited fine-tuning data, and the optimal number of shots is closer to one. At the very least, for this Siamese Neural Network, One-shot, seems to be the most effective version of “Few-shot Learning”.