**A Hierarchical Transformation-Discriminating Generative Model for Few Shot Anomaly Detection**

Anomaly detection is the process of recognizing anomalous samples in data. A large collection of training samples is frequently necessary for anomaly detection. Given a large number of samples from the normal (non-anomalous) data, the learner is typically asked to categorize novel samples as either normal or anomalous in the context of one class classification. With only a few training samples available, including the case of one training sample, the goal in this study is to overcome this problem. The lack of training data in many visual domains, as well as the human capacity to complete this task after watching a very small number of samples, are the driving forces behind our research. Reconstruction, classification, and distribution-based methods are the three categories into which image-based anomaly detection techniques can be separated. Normal data can be recreated with little error using reconstruction-based methods, but anomalous data has a large reconstruction error. Methods based on classification try to distinguish between regions of normal and anomalous data. Recent self-supervised algorithms aim to identify a "proxy" classification target that, when used to classify normal data, allows for a clear distinction between normal and anomalous data. Models based on distributions simulate the distribution of normal data.

In the proposed method, images residing the set of all natural images, are considered anomalous. The main task is to learn a classifier from the training set such that classifier detects the images in the set of all images. A multiclass discriminator is used to distinguish between different classes of transformations applied on real and generated samples. Bicubic down sampling is used here. Patch GAN is used to create samples. Each patch-GAN consists of a fully convolutional generator and discriminator. The size of receptive field is 11×11. A predetermined set of M differentiable transformations is applied to all input images, both real and generated. The transformations are chosen in a way that, on the one hand, it enriches the sample space that the current model uses to represent the distribution of images, making that model more accurate. However, the samples outside of the class are not created. The generator and the discriminator each consist of five convolutional blocks. Each convolution block consists of a 3 × 3 convolutional layer and a batch normalization layer, along with padding. Padding is done so that the spatial resolution of the input is maintained. The activation function used is LeakyReLu. When five of these blocks are used, the generator and discriminator's effective receptive fields are fixed at 11x11 for each scale. In the last convolution block, tanh is used instead of LeakyRelu. The Adam optimizer is used with the learning rate of 0.0005. The hyperparameters β1 and β2 is 0.5 and 0.999.

The proposed model makes use of both distribution- and classification-based components, which has the following advantages: A scale-dependent discriminator, which learns to discriminate between real and fake picture patches as well as between transformations of such patches, and a hierarchical generative model used to describe the internal multi-scale patch distribution of a single or few images. These elements are incorporated into the model to create a single model. In contrast to earlier methods, this model enables the detection of anomalous samples even when just one or a small number of images from the normal class are provided during training. In the few-shot condition, the proposed model performs significantly worse than the best algorithms currently available. It does not scale well to hundreds or thousands of training images without further adjustments, and training becomes more complicated as the number of training images rises. Given the emphasis in the current literature on modeling the form of the variability between the training samples, the method's one-class classification abilities in the case of a small training set, for which it was created, are surprising.

**CpG Transformer for imputation of single-cell methylomes**

During DNA methylation, a methyl group is added to the DNA. The most well-known type is CpG methylation, in which CG dinucleotides have a methyl group added to the C-5 position. There are various techniques available for measuring DNA methylation at the single-cell level. These techniques make use of DNA bisulfite conversion, followed by genome-wide and reduced-representation sequencing. Profiling single cells presents several difficulties that are not present in bulk sequencing research because of the lesser amount of genetic material that is present in each cell. Less reads are used to cover profiled sites, resulting in noisy DNA methylation readings. The full potential of single-cell methylome analysis must therefore be utilized through the use of efficient imputation and denoising techniques. Recent advancements in self-supervised learning of natural language, BERT model provide inspiration for this study. By combining axial attention with sliding window self-attention, CpG Transformer is proposed. It is an adaption of the transformer neural network architecture to operate on partially observed methylation matrices, achieving state-of-the-art imputation performances on a variety of datasets.

CpG Transformer formulates its inputs to the transformer layers using collaborative filtering methods. The CpG matrix, coupled with the CpGs' specific locations on the genome and the DNA sequences around them, serve as the inputs to the CpG Transformer. The model receives cell identity via learned cell embeddings. Every CpG site is given a representation by the model, which then integrates the data in a graph-like structure. The input to model is a three-dimensional tensor. Rows represent the cells; columns represent the methylation site of the methylation matrix. Three-dimensional embedding is produced by concatenating three embeddings in a linear fashion. To generate outputs for each input entry in a matrix, the input values are normalized and multiplied by a value matrix that was created by linearly combining the input with learning weights. Row-wise sliding window self-attention operation is provided with relative sinusoidal positional encodings to transmit the relative distances of CpG sites to the model. The model uses a stack of four identical layers. Three sublayers make up each layer. The first and second sublayers are made up of the column wise sliding window self-attention row-wise sliding window self-attention. Each has 8 heads and 8 hidden dimensions. The last sublayer uses a feedforward network with two linear combinations that is position-wise fully connected and a ReLU activation. Around all sublayers, a residual connection and layer normalization are used. To produce final predictions for all inputs, the outputs of the final transformer layer are reduced to one hidden dimension by an output head and placed through a sigmoid operation. MLM objective is adapted for DNA methylation imputation. MLM is a form of denoising autoencoding where the loss function only affects the subset of disrupted inputs.

CpG Transformer can enhance single-cell methylation studies greatly. Combining axial attention with sliding window attention, CpG Transformer adapts the transformer architecture to operate directly on methylation matrices, offering a general-purpose method of learning relationships between nearby CpG sites both within- and between-cells. Over DeepCpG, cutting-edge imputation performances are attained using CpG Transformers. Second, proposed approach is amenable to interpretation and transfer learning. Finally, CpG Transformer is expected to scale better to future larger datasets comprising a variety of cell types since its model architecture employs learned cell embeddings to encode cell identity in a flexible manner. To enable inputs with many cells, more expansions of the proposed axial attention could be made. The model may be unable to accurately estimate local methylation profiles in situations with severe sparsity, which is another drawback to this approach. The CpG Transformer lacks the ability to handle systematic noise and missingness. In these circumstances, models will most likely spread and increase the noise, potentially impairing the outcomes that are biologically important.