**Revisiting Contrastive Learning through the Lens of Neighborhood Component Analysis: An Integrated Framework**

Contrastive learning has received a lot of attention and lately emerged as one of the most successful representation learning strategies. Due to instance-level categorization, contrastive learning creates an embedding aim based on the presumption of semantic similarity between positive pairs and dissimilarity between negative pairs. In the existing approaches, each individual image in a dataset was treated to belong to its class and multi-class classification was done under this setting. Memory bank was designed for storing the representations and use noise contrastive estimations. However, all these methods have their own drawbacks in terms of computational costs and robustness. A novel stochastic nearest neighbor viewpoint of contrastive learning is provided here. A series of contrastive losses is proposed that outperform the existing ones by achieving robustness and good accuracy on downstream tasks.

NCA is a supervised machine learning algorithm that deals with learning a quadratic distance with a matrix such that the performance of nearest neighbor classification is maximized. The connection of NCA to the contrastive loss can be established by assuming that the positive pairs belong to the same class and the transformation of the of the matrix is parameterized by a general function. The reduction of NCA formulation to the contrastive loss assumes that the expected relative density of the positives is 1/N (N is the data batch size) in the underlying data distribution and the probability induced by the encoder network is 1. LNCA is obtained by relaxing the expected relative density of the positives. This is a more general contrastive loss which includes more than one element or equivalent to the number of positive pairs. LMIXNCA is obtained by allowing the neighborhood samples that agree with each other’s probability to relax. IntNaCl framework is proposed which not only has good accuracy and robustness but also generalizes the existing methods. IntCl is an intermediate integrated loss that consists of two components – standard loss and robustness-promoting loss. The LNaCl is used to strengthen the standard loss in LIntCL. This is done to design a generic loss that accounts for the robust accuracy while keeping the clean accuracy.The three major properties of representation learning methods are standard discriminative power, transferability, and adversarial robustness. Evaluating the performance difference in IntNaCl framework when the positive pairs is greater than or equal to 1 can be done by evaluating the effect of LNCA. When α = 0 and the positive pairs is not equal to 1, LNCA and LMIXNCA can both improve in all metrics (standard, robust, transfer accuracy). When α = 1, the networks show a high transfer standard accuracy, which proves the ability of learning representation networks that transfer robustness property. Evaluating the performance difference in IntNaCl framework can be done by evaluating the effect of LRobust. When the positive pairs is equal to 1, the robust and the transfer accuracy increases. When the positive pairs are not equal to 1, there is a jump in the robust accuracy.

The contribution of this work is the fact that the relationship between contrastive learning and NCA is established, and two new contrastive losses dubbed as NaCl (Neighborhood analysis Contrastive Loss) are proposed. A generic framework called Integrated contrastive learning is proposed. It can also be seen that the IntNaCl is very effective in improving the standard and robust accuracy.

**scBasset: Sequence-based modeling of single cell ATAC-seq using convolutional neural networks**

Single cell ATAC-seq (scATAC) holds enormous promise for examining cellular heterogeneity in epigenetic landscapes, but because to its high dimensionality and sparsity, it still presents substantial processing hurdles. In this paper, scBasset is introduced which can be used to model scATAC data. PCA analysis and LSI perform the linear transformation to project the cells in a low-dimensional space and are sequence free methods that can detect the biological covariance to represent and cluster the cells. But the sequence information is ignored and post-hoc motif matching tools are used to relate accessibility to transcription factors. In sequence-dependent models (Ex: chromVAR) the peaks are represented by the TF motif or k-mer content and these features are aggregated across the peaks to learn the cell representations. But, these tend to perform worst because, of the information loss from the model relating sequence. In this paper, a sequence-dependent model based mainly on deep CNNs is proposed. The Basset architecture is extended to predict the single cell chromatin accessibility from sequences.

The scBasset predicts the binary accessibility vectors for each peak based on its DNA sequence. This is an extension of the Basset architecture. A bottleneck layer is used to learn low dimensional representations of the cells. scBasset takes 1244 bp DNA as the input from each peak center. One-hot encoding of the input is done which results in 1344x4 matrix. It consists of 1D convolution layer of size 17x4 with 288 filters followed by batch normalization, Gaussian error linear unit (GELU), and width 3 max pooling 420 layers. This generates a 488×288 output matrix. It consists of Convolution tower of 6 convolution blocks each consisting of convolution, batch normalization, max pooling, and GELU layers. The convolution layers have increasing numbers of filters and kernel width 5. The convolution tower gives 7×512 matrix as the output. A 1D convolution layer with 256 filters with kernel width 1, followed by batch normalization and GELU. The output of this is a 7×256 matrix. This is flattened into a 1×1792 vector. The dense bottleneck layer consists of 32 units and followed by batch normalization, with dropout (rate=0.2), and GELU. A compact peak representation vector of size 1×32 is the output. A dense layer is used for predicting the continuous accessibility logits for the peaks in every cell. Final sigmoid activation is used for the accessibility probability. The scBasset predicts single cell chromatin accessibility on held-out peaks. The scBasset is evaluated on the additional scATAC datasets. The denoising evaluation is based on the hypothesis that in the multiome measurements, the effective denoising will improve correlation between accessibility at genes’ promoters and expression. The denoising performance is evaluated by computing the Pearson correlation between the gene accessibility score and gene expression across all genes for each individual cell.

The scBasset is trained to predict individual cell accessibility from the DNA sequence underlying ATAC peaks. As part of the training process, it learns a vector embedding to represent the individual cells. The model learns to recognize the TF motifs and the influence they have on the accessibility. When compared to the previous methods, the scBasset has a better performance in learning all the cell embeddings and the interfering TF activity. The scBasset is a more interpretable model that can be directly queried for identifying regulatory sequences or for TF activity. However, the memory efficiency of the scBasset model can be improved to scale very large datasets by sampling mini batches of both the cells and sequence. The ATAC analysis mainly depends on accurate peak calls, this can be further improved by taking the sequence information and accounting for Tn5 transposition bias into consideration.