Cooper Ang (20768006)

SYDE 572 Project Proposal

Deep k-Means architecture changes for better performance and/or Semi-Supervised Learning with Deep k-Means

Introduction

Traditional clustering methods such as k-means and Gaussian Mixture Models (GMM) rely on the provided embedding space or a lower dimensional embedding space through a linear transformation such as PCA. This is especially difficult in the domains of image and text, where many of the inputs may be sparse and in high dimensionality. A potentially better approach is to use an autoencoder to learn the representations of the data in a reduced embedding space, and apply a clustering algorithm such as k-means to find suitable clusters. Deep k-Means (Fard et al., 2020) accomplishes this by making clustering and learning the representations in the lower dimensional space jointly, through a shared optimization through a continuous reparametrization of the objective function. This optimization is done with stochastic gradient descent.

Novel Contributions

As the project is in the early stages two avenues for novel contributions are being investigated and are subject to change. Most likely only one will be completed depending on the feasibility, ease of implementation and results. An initial investigation by literature review has been conducted but there are no current implementations or results.

Architecture Changes

The authors of the Deep k-Means provided a default architecture for the multilayer perceptron that acts as an encoder (d-500-500-2000-K), where d is the input dimensionality and K is the number of clusters. The decoder is a reflection of the encoder. There is room for improvement here by potentially using a convolutional autoencoder, which has in cases such as Medical Image Analysis has done a better job at deep feature learning for supervised and unsupervised learning (Chen et al., 2021). The chosen convolutional architecture of the autoencoder to be used with the Deep k-means algorithm will need to be empirically verified, and may increase or decrease performance. An image dataset such as Fashion-MNIST (which is not used in the paper) can be used to verify baseline MLP autoencoder Deep k-means versus the convolutional autoencoder Deep k-means variant. Multiple MLP architectures with different numbers of layers and weights can also be experimented with to attempt to beat the referenced design in the paper.

Semi-Supervised Learning

There are multiple augmentations proposed to the traditional k-means algorithm that constrain the clusters with a few labeled examples. This is termed “semi-supervised

learning", as the dataset contains a mix of a small amount of labeled data and a large amount of unlabeled data. The idea here would be to apply an algorithm such as the PCKmeans (Basu et al., 2004), which adds a penalty term to the standard K-means criterion to minimize the amount of points not satisfying the constraints, and integrating this constrained approach to the selection of clusters for the Deep k-means algorithm. The difficult part may be how that would work from the mathematical perspective, especially how it would change the gradients used to optimize the autoencoder and the clusters. This may or may not be feasible and more study must be done to understand the potential novel contribution.

Motivation

Deep neural networks are becoming popular in pattern recognition, as the compute and the amount of data increases. It is an important problem space to investigate the connection between traditional pattern recognition algorithms such as (K-means) and deep learning techniques such as autoencoders, and convolutional neural networks. Dimensionality reduction and dealing with sparse data are some important problems currently facing the pattern recognition field, especially as modalities such as NLP and computer vision are becoming popular.

Problems related to unsupervised and semi-supervised pattern recognition are important because data is expensive to label, and without a fully labeled dataset there is no choice but to use these types of algorithms. There is a lot of potential in the applications of combining deep neural network techniques and traditional methods such as k-means to accomplish unsupervised and semi-supervised learning.

From a personal learning perspective it will be an interesting experiment to learn and understand how the traditional loss function of a neural network is augmented to include the K-means algorithm. Having mostly had experience with vanilla MLPs and Transformers in the domain of NLP, I hope to gain more understanding of the mathematics behind optimization and working with image datasets in the computer vision domain. I also don't have any experience with the "reparameterization trick" in order to make the function differentiable, which is seeing wide use in new models such as diffusion based models, so through this project I hope to gain a better understanding of this, and expand from the confines of traditional pattern recognition techniques.

References

Basu, S., Banerjee, A., & Mooney, R. J. (2004). Active semi-supervision for pairwise constrained clustering. *Proceedings of the 2004 SIAM International Conference on Data Mining*. https://doi.org/10.1137/1.9781611972740.31

Chen, M., Shi, X., Zhang, Y., Wu, D., & Guizani, M. (2021). Deep Feature Learning for medical image analysis with convolutional Autoencoder Neural Network. *IEEE Transactions on Big Data*, *7*(4), 750–758. https://doi.org/10.1109/tbdata.2017.2717439

Moradi Fard, M., Thonet, T., & Gaussier, E. (2020). Deep K-means: Jointly clustering with K-means and learning representations. *Pattern Recognition Letters*, *138*, 185–192. https://doi.org/10.1016/j.patrec.2020.07.028