Self-Supervised Learning for DeepLense Implementation

Philip LaDuca

philip_laduca@brown.edu

Summary:

Self-supervised learning has promising applications to gravitational lensing due to the limitations regarding data distribution and labels. This project aims to add a variety of self-supervised learning models to DeepLense for both classification and regression tasks. The models will include both contrastive and clustering self-supervised learning methods as well as explore hybrid and equivariant transformer architectures in a survey of contemporary methods.

Overview:

Gravitational lensing images commonly lack labels or uniform distributions across classifications. While these limitations make supervised learning a challenge, self-supervised learning has been shown to outperform supervised models in these limited data regimes. Self-supervised learning enables models to learn from unlabeled data by leveraging certain features or relationships in the data to create their own labels. This allows the self-supervised models to learn representations of the data in the form of embeddings, capturing underlying structure. With these representations, the model is able to be fine tuned on much smaller labeled datasets for tasks such as regression or classification while not being hindered by the small amount of data.

Recently, transformers have found applications in computer vision with the introduction of the vision transformer. This has led to a plethora of models being created with varying strengths and weaknesses. In a natural progression, self-supervised learning techniques have begun to be applied to vision transformer models to push their capabilities even further. However, when applying self-supervised learning, vision transformers are prone to instability which leads to

representation collapse (Ling et al, 2022). Fortunately, these issues can be fixed when convolutions are added to the patch projection layer. This leads to the combination of hybrid architectures and self-supervised learning as a contemporary approach to classification and regression models. Furthermore, equivariant transformers show promise due to the geometric nature of gravitational lensing data. As such, equivariant architecture will also be explored in a self-supervised context.

Project:

In order to find the best architecture to use for gravitational lensing data, it is important to survey a variety of different models and compare the results. When building a self-supervised model, there are two main components, the self-supervised method and the model architecture. Normally, self-supervised methods are modular, allowing for application to a variety of models. Therefore, it is important to test each self-supervised method with a variety of model architectures. For this survey, I will be choosing 7 different self-supervised methods as well as 3 different architectures resulting in 21 different models to be trained. The methods and architectures are picked such that they are representative of the contemporary strategies being implemented. In order to test the ability of these models, two types of datasets will be used: one with few labels and the other with a skewed distribution (eg. variable number of data points in each class or a non-uniform mass distribution). The model can be trained with self-supervised learning then fine tuned in different manners for either classification or regression to test the transferability of the model architectures.

Self-Supervised Learning Methods:

- SimCLR (Chen et al, 2020a): A contrastive self-supervised method that leverages both positive and negative examples in a batch. The model minimizes the distance in

embedding space between positive pairs while maximizing the distance between negative pairs. This is a good example of a simple contrastive method.

- SimSiam (Chen et al, 2020b): A contrastive self-supervised method that focuses only on positive pairings by comparing two augmented versions of the same image. In order to prevent collapse, a stop gradient is added. This is a simplified version of SimCLR which still obtains competitive results.
- SwAV (Caron et al, 2020): This approach utilizes both clustering and contrastive learning by training the model to predict whether two views of an image belong in the same cluster or not. It uses strategies such as prototype vectors and swapping of assignments to learn more robust representations. It is a step away from the previous Sim networks towards more cluster based self-supervised learning.
- MoBY (Xie et al, 2021): This method is the combination of two other self-supervised learning methods MoCo and BYOL. MoBY takes the momentum and queue aspects of MoCo and combines it with the asymmetric encoders and momentum scheduling from BYOL creating a dual encoder for contrastive learning where the second encoder is updated based on an exponential moving average of the first. The queue is used as a dictionary where the query is matched to its target feature and separated from the others. This approach utilizes the queue architecture not found in other methods.
- DINO (Caron et al, 2021): This method uses a dual network known as student and teacher using momentum once again to update weights of the teacher. The models are also asymmetric with the teacher applying a centering to the representation and standard cross entropy loss being used for contrastive learning. This is a second momentum architecture this time incorporating some clustering due to the manner in which the cross entropy loss impacts the learning.
- SiT (Atito et al, 2021): This method utilizes both reconstructive learning as well as contrastive learning. A transformer is given two corrupted versions of images and is

tasked with reconstructing the original. The resulting loss from the reconstructive task is then combined with the contrastive loss between the representations to update the parameters of the network. This model introduces reconstructive learning as a strategy for self-supervised learning.

- VTCC (Ling et al, 2022): This method utilizes two different projectors to take the representations from the visual transformer and calculate instance and cluster loss. Instance loss is a contrastive loss between the two different inputs while the cluster loss is a contrastive loss focused on clustering in the dataset. This model combines clustering and contrastive learning in a direct way.

Transformer Architectures:

- CvT (Wu et al, 2021): This hybrid model introduces convolution into both the token embedding as well as the transformer block of a vision transformer. Prior to creating patches, a convolution is passed over the image reducing the token taken in by each transformer block. Then, in the transformer block, convolutional projections are passed into the attention layer. Both of these strategies are used to enrich the representations created by the model through the convolutions helping capture structure in the image.
- LeVit (Graham et al, 2021): This hybrid model introduces convolutions prior to the transformer blocks but also leverages shrinking attention blocks to create a pyramid structure. Similar to many convolutional networks with pooling, as the representation passes through the model the shrinking attention layers reduce the resolution of the activation maps.
- Equivariant Transformer (Tai et al, 2019): The equivariant transformer differs from the other transformers since they are constructed to be equivariant to different transformations. This is done through various layers which utilize mathematical symmetries to help the transformer stay consistent despite transformations being applied

to the input. This helps the model capture patterns and relationships in the data that are invariant to the symmetries in the problem domain.

Background:

Prior to working on this project I have had experience working with both gravitational lensing data as well as machine learning. I worked as an undergraduate researcher for Professor Ian Dell'Antonio at Brown University doing image processing on weak lensing images utilizing his image pipeline. This allowed me to familiarize myself with gravitational lensing as well as gain experience handling data. I also completed a project as a part of his group finding a statistical luminosity function for galaxy clusters and while this was not directly working with lensing, it gave me ample experience working with image data. I have also worked as an undergraduate researcher for the CRUNCH group at Brown focused on physics informed neural networks(PINNs). This gave me experience creating neural networks from scratch in JAX as well as reading and understanding academic papers. On multiple occasions I was tasked with taking a paper and implementing the PINN it showcased for use on a different set of equations. I was even able to fully implement the Separable PINNs (Cho et al, 2022) paper on my own in two weeks without the code having been released. My background in both astrophysics and computer science has prepared me to not only efficiently implement contemporary machine learning algorithms but also leverage the scientific background of gravitational lensing to excel on this project.

References:

Atito, S., Awais, M., & Kittler, J. (2022). SiT: Self-supervised vIsion Transformer. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2104.03602

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2021). Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2006.09882

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2104.14294

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020a). A Simple Framework for Contrastive Learning of Visual Representations. ArXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/2002.05709

Chen, X., & He, K. (2020b). Exploring Simple Siamese Representation Learning. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2011.10566

Cho, J., Nam, S., Yang, H., Yun, S.-B., Hong, Y., & Park, E. (2022). Separable PINN: Mitigating the Curse of Dimensionality in Physics-Informed Neural Networks. doi:10.48550/ARXIV.2211.08761

Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., & Douze, M. (2021). LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2104.01136

Ling, H.-B., Zhu, B., Huang, D., Chen, D.-H., Wang, C.-D., & Lai, J.-H. (2022). Vision Transformer for Contrastive Clustering. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2206.12925

Tai, K. S., Bailis, P., & Valiant, G. (2019). Equivariant Transformer Networks. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/1901.11399

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). CvT: Introducing Convolutions to Vision Transformers. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2103.15808

Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., & Hu, H. (2021). Self-Supervised Learning with Swin Transformers. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2105.04553