Semi-Supervised Deep Learning for Molecular Structures

Alvin Wan and Allen Guo
Third-Year Undergraduate Students

Introduction

Recent advances in microscopy make possible 3D nanometer-scale imaging of biomolecules. However, interpreting these images typically requires manual labelling of molecular structures using 2D projections of the data.

Our dataset contains 3D images of clathrin, a protein involved in cell transport. During clathrin-mediated endocytosis (CME), clathrin surrounds the molecules to be transported, forming a spherical coat. We applied semi-supervised learning to identify images of clathrin undergoing CME.

Our data was provided by the Ke Xu lab in the College of Chemistry, whose research work we are supporting.

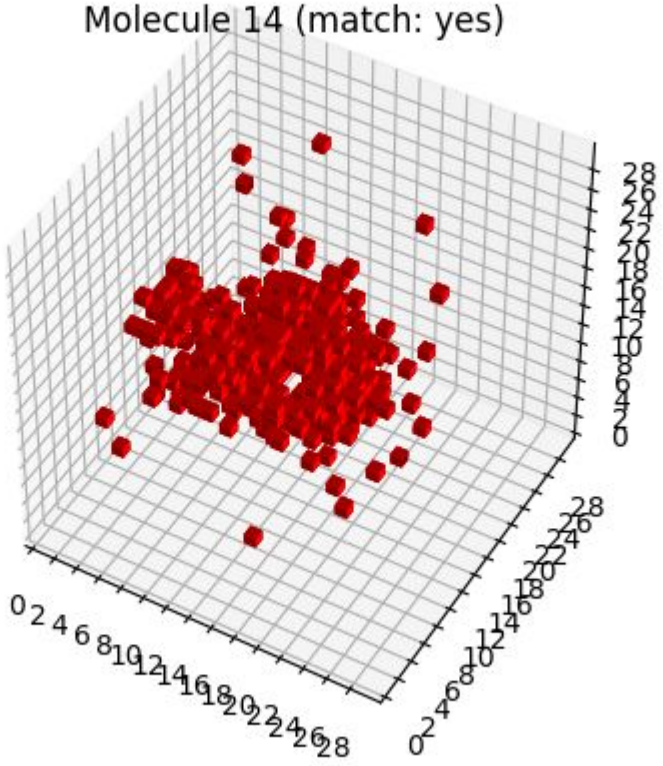
Related Work

Most existing work is tailored to urban settings and indoor environments, whereas this project is focused on objects with potentially many orientations and related structures.

One proposed speedup found in the literature considers reducing the 3D classification problem to multiple 2D projections, to save time. However, microscopies have highest variance in only one hyperplane, minimizing the efficacy of this approach. In short, previous solutions are insufficient for challenges specific to microscopies and molecular data.

STORM Dataset

Our clathrin dataset comprises of 280 labeled images and 619 unlabeled images. Each “image” is actually a *point cloud* (a set of points in 3D space), which we downsampled by dividing the point space into equally sized cubes (*voxels*) and identifying cubes that contain points. The result is a 30x30x30 binary 3D image for each example in our dataset.



Approach

Our goal was to train a classifier that could identify whether a given clathrin image shows clathrin involved in CME or not.

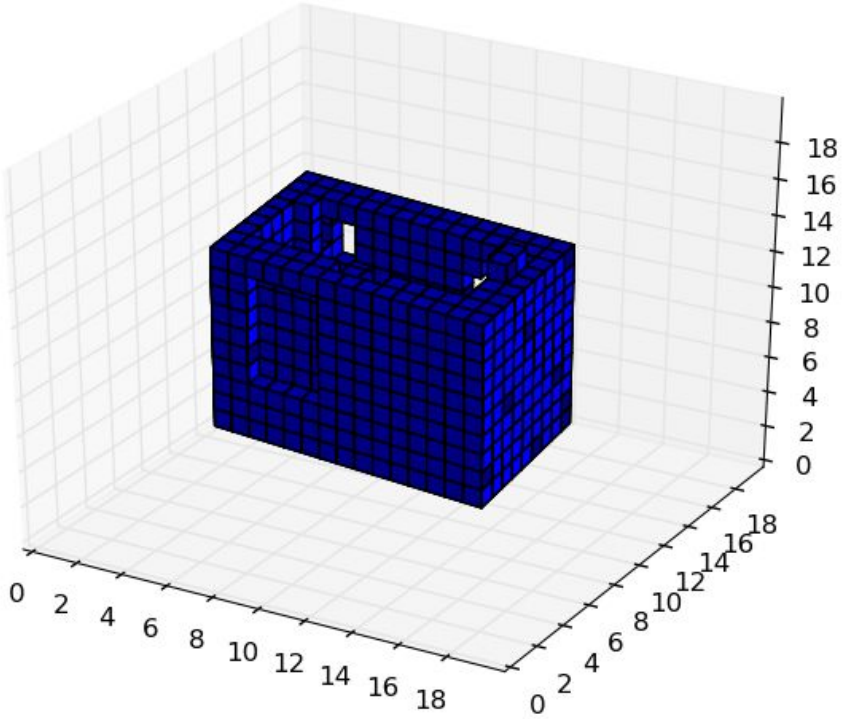
Because our dataset is so small, we aimed to leverage the relatively large amount of unlabeled data to perform *semi-supervised learning*: learning good featurizations of the input data using the combined labeled/unlabeled data, then training a classifier using only the featurized labeled data.

We tried the following unsupervised approaches: feedforward autoencoders (FAEs), convolutional autoencoders (CAEs), PCA, and soft *k*-means. On top of the featurized data, we trained an SVM to perform classification.

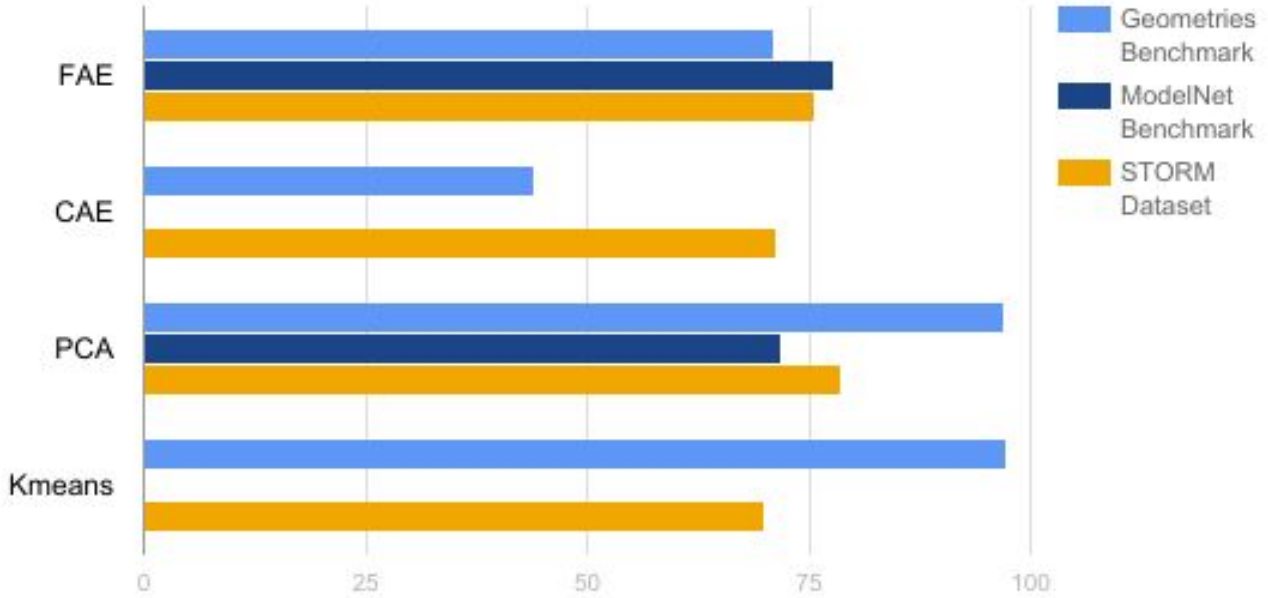
Benchmark Datasets

We validated our approach on two additional 3D image datasets:

- **Geometries**, a homegrown dataset containing 600 geometric figures (cubes, spheres, and diamonds) with additive noise.
- **Princeton ModelNet10**, which contains ~57,000 real-world objects drawn from 10 categories (e.g., bathtub, on right).

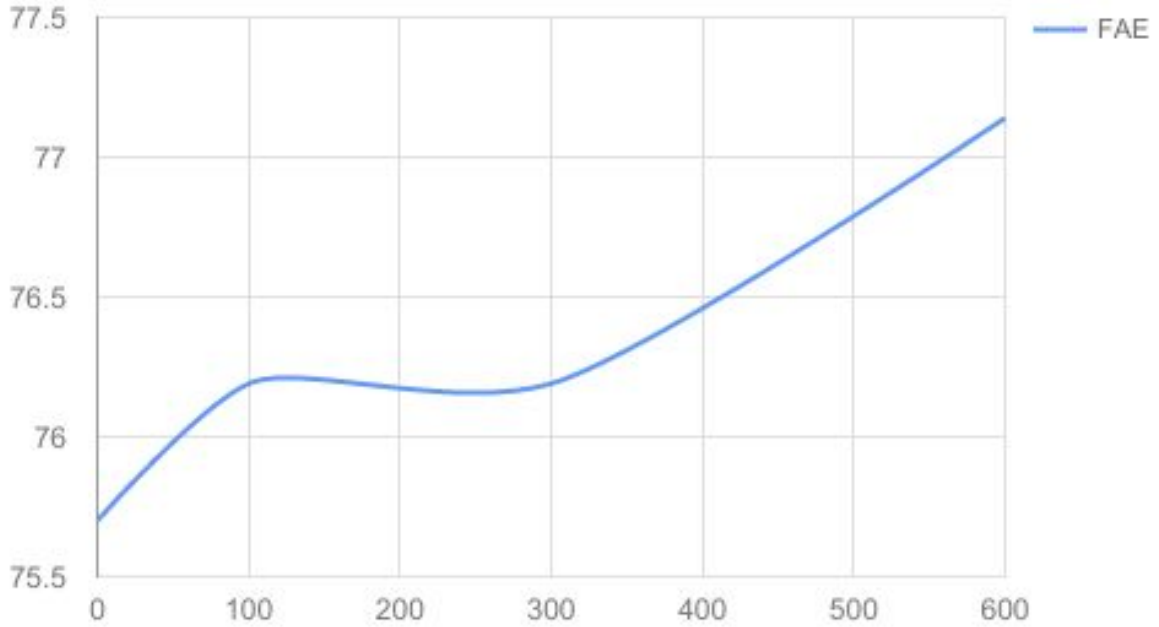


Semi-Supervised Performance of Various Encodings



All encodings used 64 latent dimensions. The highest accuracy (not listed) was PCA with SVM for 100 latent dimensions at 78.57%. Kmeans did not finish running for ModelNet, and the CAE ran into a memory error. Both neural networks were ran for 10 epochs. We used a grid search for the optimal SVM hyperparameters. We noted **reconstruction accuracy was not correlated with classification accuracy**.

Performance Using Unlabeled Data for Encoding



The fully connected autoencoder saw gradual improvement with an increasing number of unlabeled samples. The convolutional autoencoder saw measly performance, at a maximum of 61.43% and minimum of 60.0%.

Conclusion

Our best unsupervised learning method, PCA, performed respectably on the clathrin dataset, achieving nearly 80% accuracy.

Conversely, deep neural networks excelled on the ModelNet benchmark dataset, but did not perform well on the clathrin dataset. We believe this is because there is simply not enough data to train on.

Ultimately, we expect that our work will enable the Xu lab to classify clathrin and other molecular structures more efficiently.

By our final report, we hope to 1) try deep learning with data augmentation and 2) obtain a subjective evaluation of our work from the Xu lab.