Image Clustering Over Time

Dylan Lee, Zach McMichael, Viktor Zenkov, Adam Tutko

Abstract— **Image analysis has become a large part of deep learning and machine learning research. In this work we will apply some of these deep learning methods to generate image representations which will then be clustered by some generic clustering algorithms, such as K-Means. The goal of this will be creating clear and concise clusters that track the same feature in an image, even when that feature shows change over time. With successful results this will further the work currently being done on human decomposition images and show that these algorithms can be generically applied to any similar data set or domain.**

I. Objective

The goal of this project is to apply deep learning and clustering algorithms to create groupings of images that track the same feature over time. One of the main objectives related to creating these clusters and analysis will be cultivating or discovering an appropriate image data set that tracks a feature through time, and preferably is labeled and contains some type of metadata. Once a data set is found, then the goal will be applying the already developed deep learning model and clustering methods to determine how successful and accurate these methods can cluster other image sets from different domains.

This project will further develop the deep learning model used to generate image representations and apply similar methods as used in the ICPUTRD[1] research to take advantage of metadata to improve cluster accuracy. This includes using the dates of when the images were taken with the goal of merging clusters between days to improve on feature tracking over time. The data discovery aspect of this project will be one of the most important and consist of finding an appropriate data set, collecting and organizing the metadata. Overall, this project will dip into all the focus areas of the class with a direct focus on data discovery and analysis.

II. Motivation

The motivation behind this project is related to work currently being done on an image data set from the body farm that tracks human decomposition (ICPUTRD). While the results from this work have shown promise, this project will test the algorithms and methods to see how well they work in other domains on different data sets. This project was also chosen due to the interest in deep learning and CNN's currently being used to create the image representation. Another motivating factor is the challenge of finding an appropriate data set for the project. While there are many labeled data sets available, finding one that has images of the same feature, where that feature shows changes over time, is a much more difficult task. The final motivation of the team was to choose a project in which our work will benefit and enhance other research being done in this field. Since a team member is already involved in research on this topic, there is a natural easy overlap where this project will hopefully benefit the research and vice-versa.

III. Team

The team will be made up of four senior Computer Science majors that all have experience in machine learning but little experience in image analysis and CNN's. Zach McMichael has worked with many machine learning libraries while doing text analysis on tweets and is experienced in C++, Python, and Java. Dylan Lee has worked on the ICPUTRD research team and gotten experience with using the deep learning models and modifying clustering algorithms to take advantage of metadata, he is experienced with C++, Python, and Java. Adam Tutko has experience with text analysis on open source git repositories, and has worked in C++, Python, and Java. Viktor Zenkov has experience with machine learning for text analysis and regression and has experience with C++ and Python.

IV. Data

A very important part of this project is data collection. Since the algorithms that have already been developed and tested are centered around a specific data set, it is a difficult challenge to find a large collection of images that also track a feature that changes over time. Some early thoughts of the team were finding image sets that track different types of mushrooms or plants growing, or even using a data set from NASA tracking constellation and planet movement. However, the most promising data set that was found is an image data set named MORPH that is a collection of "mug-shot" photos that track different subjects over time. This image data set contains thousands of different individuals with multiple photos ranging from a few days to a few years apart. This provides a similar challenge to the body farm images, where the goal is to track features that changes over time; the face is the feature and the effects of aging over time take the place of decomposition. This data set also provides a good amount of metadata such as age, race, and date that can be used to improve the clustering. The downside to this data set is it is not open source, and is only available to license for commercial use or purchase for academic and research use. Another issue is that this image set only shows one feature, in this case, a person's face. While this is not as unique and vast a data set as the ICPUTRD images, it still provides the

[1] ICPUTRD is a research project focusing on clustering and grouping human decomposition images using deep learning and CNN repersentations.

best progression of a feature over time that the team was able to find. To obtain access to this data set it was purchased for $99.99 from The University of North Carolina Wilmington. Once obtained the metadata and image paths were organized and collected into a mongodb to provide easier retrieval. The next step that was taken to "clean" the data was narrowing down to only the individuals that have more than 15 images a piece. This was necessary to remove individuals that only contain a few images, as testing on these individuals does not match the original goal of the algorithm. The goal is to build clusters that track a sequence of changing features, not just classify individual photos. Once the data set was condensed down to high image count individuals the team was left with around 5000 images of 300+ individuals. While this still isn't as vast and varied of a data set as ICPUTRD it still provides a substantial test for the algorithm in question.

V. Methods

This project utilized two different deep learning models already implemented in Python3, ResNet50 and VGG16. Both of these models were used already pretrained on ImageNet, a large labeled dataset with over 1000 classes. To establish a baseline precision for clustering this dataset, ResNet50 and VGG16 were used to generate feature vectors for each image in our dataset and these were saved in our mongodb. The feature vectors were then used with two difference clustering algorithms, KMeans and Agglomerative, and the results analyzed. To visualize and improve the precision of this baseline method, PCA and TSNE dimension reduction was preformed and the final baseline was chosen as the more accurate of these two, which resulted in the team using TSNE dimension reduction. Once the baseline precision was established the original full length feature vectors were modified by appending metadata weights such as race, gender, facial hair, glasses, etc. These weight were normalized with the other values in the feature vector. These new vectors were then reduced using the same TSNE dimension reduction as in the baseline method and clustered using KMeans and Agglomerative algorithms. Other tools used in the evaluation of these clusters and feature vectors included an open source web based visualization tool for clustering called parallax, developed by Uber research, and the final clustered were plotted using Python's Matplotlib.

VI. Final Results

The final results for this project matched very closely to the expected results. The baseline methods showed an average precision of roughly 60% with a higher precision for the agglomerative clustering. This is a reasoable baseline to expect with a pretrained model, and with the appended metadata the team expected roughly a 20% increase in the precision. After the modifications the precision of the clusters increased by a full 30%. The final precision of the models is shown in the table below.

Model	KMeans	Agglomerative
Baseline VGG16	59.832%	67.728%
Baseline ResNet50	63.286%	69.759%
Modified VGG16	98.035%	99.205%
Modified ResNet50	97.628%	99.192%

VII. Issues

Overall this project went relatively smoothly and no major issue were encountered. There were a few smaller issues, however, that had to be overcome. The first of these was modifying the file structure of the data set to make it more accessible and to keep the paths to the images consistent. This was easily done by writing a few bash scripts to reorganize the images. The second hurdle that had to be overcome was ensuring all members of the group had access to the images without making them publicly available. This was accomplished by creating a mongodb docker container that was accessible for all team members. The last issue the team dealt with was ensuring that the algorithm was only tested on individuals that had more than 15 pictures in the dataset. Again, this was accomplished by writing some bash shell scripts to count the number of photos for each individual and only gather the paths for these images rather than all 55,000+ images. None of these issues caused any major problems, and the team was happy with how smoothly this project progressed.

VIII. Future Work

Future work that would be related to this project would involve expanding the tested algorithm to other data sets gathered from different domains. Other than just applying the current algorithm to other datasets, more work could also be done on the algorithm itself and the team thinks it would be very interesting to try different deep learning models and also different clustering algorithms to evaluate the most successful. More work could also be done on actually training a deep learning model to generate feature vectors, since now a pretrained model is being used. If a new model was trained, cross-validation would be needed to ensure no over fitting, but with time the team feels like this could provide better clustering and overall better results. The most interesting future work would be turning the current algorithm into an abstracted framework that could work on any image dataset, regardless of the file structure or metadata available. This would require much more time than the team has available but is the most intriguing future work for this project.

IX. Org Chart

The responsibilities for each member.

Name	Role
Dylan Lee	Project Lead
Zach McMichael	Machine Learning Expert
Viktor Zenkov	Model Validation / Documentation
Adam Tutko	Data Set Augmentation / Documentation

The timeline for the project:

Week of	Milestone
9/30	Find or cultivate a data set
10/07	Clean and prepare image metadata for use in clustering
10/14	Generate image representations using deep learning
10/21	Using image metadata modify the clustering methods to improve cluster accuracy
10/28	Evaluate the clusters and continue to modify the algorithms
11/4	Based on results, either continue modification on algorithm or begin results analysis
11/11	Generate results analysis graph and prepare presentation of results