DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

Facial Named Entity Recognition by Attention-Based Graph Convolutional Neural Network

Garret Obenchain
Sandia National Laboratories
gjobenc@sandia.gov

Cameron Rhea
Sandia National Laboratories
cmrhea@sandia.gov

Ryan Cooper
Sandia National Laboratories
rcooper@sandia.gov

1 ABSTRACT

In the realm of facial recognition and analysis, the ability to accurately cluster large datasets of facial images stands as a cornerstone for various applications, ranging from security surveillance to user biometric identification. This project evolves a novel approach to facial data clustering by embedding facial images into a high-dimensional vector space using an advanced embedding model trained on separate data and assumes a graph-like structure on the high-dimensional vectors. We find our method works significantly better than common shallow methods.

2 INTRODUCTION

Facial recognition technology has seen significant advancements over the past decade driven by the increasing availability of large-scale datasets and the advancement of machine learning algorithms. This technology has a wide range of applications, from security and surveillance, to user authentication and social media tagging. However, one of the persistent challenges in this field is the ability to accurately cluster and recognize faces in evolving and diverse datasets. Common shallow clustering methods often struggle with the complex and high-dimensional nature of facial data, leading to suboptimal performance.

We evolve a previous approach [5] to facial named entity recognition (NER) by leveraging an attention-based graph convolutional neural network (GCN-A). Our method aims to address the limitations of conventional clustering techniques by embedding facial images into a high-dimensional vector space and assuming a graph-like structure on these vectors. This approach allows us to capture the intricate relationships between facial features and improve the accuracy of clustering and recognition tasks.

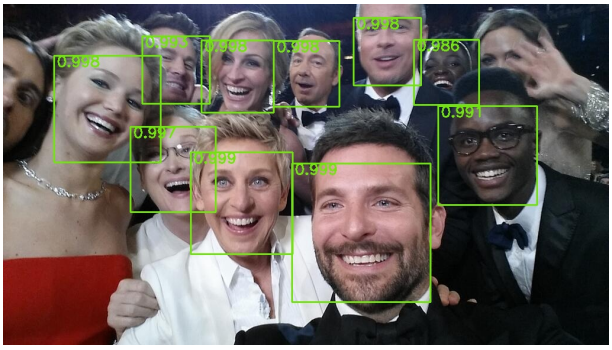

Figure 1: An example of inputs for our pipeline which are outputted from a bounding-box face detection algorithm.

3 MOTIVATION

The motivation behind this research stems from the need for more robust and adaptable facial recognition algorithms. Traditional shallow clustering methods, such as K-Means, Gaussian Mixture Models (GMM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), often make impractical assumptions about data distribution. For instance, K-Means requires clusters to be convex-shaped and the number of clusters to be specified in advance, while GMM assumes that data points are generated from a mixture of Gaussian distributions. DBSCAN, on the other hand, assumes that clusters have relatively uniform density. These assumptions can lead to poor performance when dealing with the complex and non-linear nature of facial data.

Our proposed method addresses these challenges by utilizing a graph-based approach. By embedding facial images into a high-dimensional vector space and constructing subgraphs for each embedded node, we can better capture the underlying structure of the data. The use of graph convolutional networks (GCNs) allows us to effectively propagate information across the graph and improve clustering accuracy.

4 METHODOLOGY

We present a comprehensive overview of the methodology employed in our study, detailing the processes and techniques used to achieve high-accuracy facial data clustering. The following subsections outline the datasets utilized, the embedding models selected, and the detailed steps of our pipeline.

4.1 Data

In our experiments, we utilize three widely recognized datasets: CASIA [7], DigiFace [1], and CelebA [6] where each contains widely varying distributions and sample sizes.

Table 1: The number of images and subjects in each dataset.

	CASIA [7]	DigiFace [1]	CelebA [6]
Number of Images	494,414	499,995	202,599
Number of Subjects	10,575	99,999	10,177

4.2 Preprocessing

In order to normalize the data before it is fed into the embedding models we utilized the following preprocessing.

- All faces were resized to 112×112 pixels. This standardization is necessary to maintain consistency across the datasets and to match the input requirements of the embedding models.

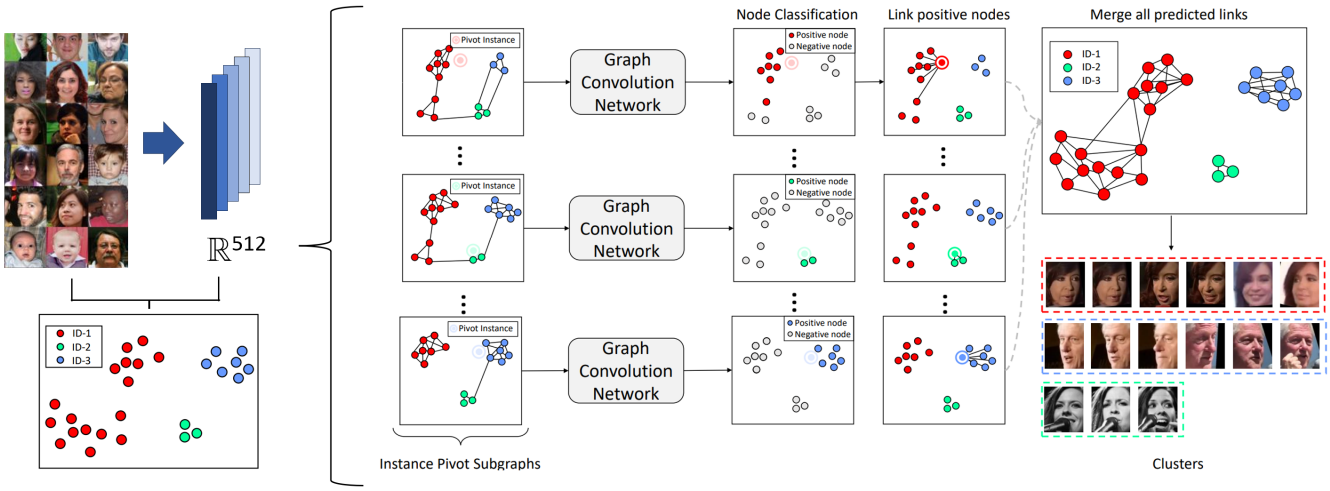

Figure 2: A visual representation of the stages of a GCN originally proposed by [5]. We embed faces into 512-dimensional vectors so they exist as points in some embedding space. Then, construct sub graphs by looking at its closest neighbors. Finally, apply graph convolution to each of the sub graphs.

- Pixel values of the images were normalized to a range of $[-1, 1]$. We performed duplicate removal to ensure that each face in the dataset is unique.

4.3 Embedding Model

Initially, three models were chosen to embed the image data into high-dimensional vector space: two models, a ResNet50 (R50) and ResNet100 (R100) trained on the MS1M-RetinaFace and WebFace datasets, respectively, and a model with an R100 backbone trained on the Glint360K dataset [2]. In addition to greater performance on multiple benchmarks reported by Guo et al. [3], the decision to ultimately use the R100 Glint360K model was made based on measures of the sparsity of embeddings. More sparsely distributed embeddings in high-dimensional vector space generally improve clustering performance since the distance between clusters is greater, making clusters easier to distinguish from one another. By measuring the pairwise Euclidean distance and cosine similarity of all the embeddings from each model, averages were calculated and showed that the R100 Glint360K model had the most sparse embeddings as seen in Table 2.

Table 2: Average pair-wise euclidean distance and cosine similarity for each embedding model

Model	Euclidean Distance	Cosine Similarity
R100 Glint360K	29.36	0.020
R50 WebFace	27.06	0.023
R50 MS1MV3	25.87	0.037

4.4 GCN

The GCN utilized in this study comprises a stack of six graph convolution layers, each activated by the ReLU function and outputs the linkage likelihood between two nodes. The objective function used for optimization was the cross-entropy loss function following the softmax activation. In practice, we backpropagate the gradient exclusively for the nodes that are immediate neighbors, focusing solely on the direct connections between a pivot and its immediate neighbors. This approach significantly accelerates the process and also enhances accuracy.

4.5 Our Pipeline

The following steps outline the detailed methodology of our pipeline and is visualized in Figure 2.

- First, we embed every face in a dataset so that it becomes a node x where every $x \in \mathbb{R}^{512}$.
- Then, we construct a subgraph SG for every embedded node x_n by finding k nearest neighbors so that x_n is the center node. We normalize the node features by subtracting x_n from every neighbor feature k_n. We then represent the topographical structure of a SG_n by making an adjacency matrix A_n for each edge in SG_n.
- Finally, given a SG and A as input data, we apply graph convolution $GCN(SG, A)$ the network outputs a set of weighted edges of the entire graph, i.e. the linkage likelihood of each corresponding edge.

5 RESULTS

The following results were acquired during our experiments.

Table 3: F1 score and Normalized Mutual Information comparison of experimented clustering methods for each dataset. Entries marked with * took too long to compute for this experiment.

	CASIA [7]		CelebA [6]		DigiFace [1]	
Method	**F1**	**NMI**	**F1**	**NMI**	**F1**	**NMI**
K-means	0.731	0.779	0.732	0.871	0.438	0.730
DBSCAN	0.253	0.194	0.710	0.688	0.057	0.060
GMM	0.717	0.768	0.882	0.953	*	*
Birch	0.791	0.816	0.928	**0.971**	0.684	0.884
GCN-M	0.966	**0.957**	**0.977**	0.942	0.842	0.872
GCN-A	**0.971**	0.954	0.963	0.949	**0.868**	**0.892**

GCN-M and GCN-A denote the aggregation methods we employed for each subgraph, representing mean-based aggregation and attention-based aggregation respectively.

6 DISCUSSION

The results of our experiments demonstrate the efficacy of our proposed attention-based graph convolutional neural network (GCN) for facial named entity recognition (NER). In this section, we delve deeper into the implications of our findings and propose some discussion points.

6.1 Adaptability and Scalability

One of the most notable advantages of our approach is its adaptability. Unlike traditional methods that require re-training when new data is introduced, our graph-based method can dynamically adapt to new labels and facial images. This makes our approach particularly suitable for real-world applications where the dataset is continuously evolving.

In terms of scalability, our method demonstrates robust performance across datasets of varying sizes and distributions. The ability to handle large-scale datasets without a significant drop in performance is crucial for practical applications such as security surveillance and biometric identification.

6.2 Limitations and Future Work

Despite the promising results, there are a few limitations to our approach that could warrant further investigation:

- **Computation:** The initial embeddings and attention-based GCN operations are computationally intensive, which may pose challenges for real-time applications. Future work could explore optimization techniques to reduce the computational overhead.
- **Generalization:** While our method shows excellent performance on facial datasets, its applicability to other domains remains to be explored and solely depends on the embedding model used. Additional embedding spaces would need to be trained to create a sparse space similar to the facial embedding space used in our method.
- **Noisy Data:** Although our method performs well on clean datasets, its robustness to noisy or low-quality data has not been thoroughly evaluated.

6.3 Real-World Applications

Our pipeline has several important implications for real-world applications. Here are some of the potential impacts and benefits of our approach in various domains:

- **Security and Surveillance:** In security and surveillance, the ability to accurately cluster and recognize faces in real-time is crucial. Our method's adaptability to new data without the need for re-training ensures that security systems could remain up-to-date with minimal downtime in a cloud computing environment.
- **Social Media and Content Management:** In social media platforms and other information platforms, the ability to automatically cluster and tag faces in images can greatly enhance user experiences or enable the ability to collect fast information about a given person.
- **Biometrics Research:** Biometric systems are increasingly incorporating multiple modalities, such as facial recognition, fingerprint scanning, and iris recognition, to enhance accuracy and security. Our method could be integrated with multimodal systems to provide a more comprehensive and reliable identification process.

7 CONCLUSION

We successfully demonstrate the effectiveness of graph-based clustering in comparison to popular shallow learners. Not only does this algorithm outperform all of our other experiments, but it also provides a means for not needing to retrain every time a new set of nodes is introduced.

8 ACKNOWLEDGEMENTS

This work was supported by Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

[1] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *2023 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023.
[2] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.
[3] Jian Guo, Jiankang Deng, Xiang An, Jack Yu, and Baris Gecer. Insightface: 2d and 3d face analysis project, 2024.
[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
[5] Zhongdao Wang, Liang Zheng, Yali Li, and Shengjin Wang. Linkage based face clustering via graph convolution network, 2019.
[6] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
[7] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch, 2014.