Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 4, 2023
A Regional Conference of the Society for Industrial and Systems Engineering

Creating a Smarter Army - The Application of Semi-Supervised Learning in Image Classification

Elijah Dabkowski, Mike Powell, and Nicholas Clark

Department of Mathematical Sciences
United States Military Academy
West Point, NY 10996

Corresponding author's Email: elijah.dabkowski@westpoint.edu

Author Note: We would like thank MAJ Nicholas Fraizer and the USASOC AI Division along with Raoul Ouedraogo and the team at MIT Lincoln Laboratory for their continuous support and interest in our research. Their insights provided new ways of thinking about problems that enabled our team to tackle these issues in a multitude of different ways. Additionally, COL Nicholas Clark and LTC Mike Powell did a tremendous job of guiding me throughout the entirety of the 2023 academic year, and their support is extremely appreciated. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Abstract: Our research demonstrates how technological and data science practices can be paired with user knowledge to both improve task performance and give the user confidence in the system they are using. In this manuscript, we focus on image classification and the problems that arise when an analyst is tasked with classifying a large amount of images in a timely and accurate manner. Utilizing a well-known unsupervised classification algorithm (k-means) and pairing this with manual classification of certain images by a user, we create a semi-supervised approach to image classification. This semi-supervised classification approach performs with a higher accuracy than a strictly unsupervised approach and takes far less time than having a user manually label every image, demonstrating that a combination of both machine and human strengths produces better outcomes faster than any alternative.

Keywords: Semi-Supervised Learning, Image Classification, Target Identification

1. Introduction

With around 500, 000 active duty personnel, the United States Army is a fighting force containing a multitude of different jobs including infantrymen, tankers, and engineers. Working to support these combat arms branches, military intelligence officers are responsible for the procurement and delivery of information that will support the Army and the individuals serving within it to ensure that those fighting on the front lines have as much information as possible (Army, 2020). After talking with those who have worked with or as an analyst in the target identification sub-field of military intelligence, there is a common idea that it is an art, not a science. We believe that current practices can be improved by science to help expedite the process of having a group of individuals sift through images or videos captured from an area in which the military is conducting operations. Consider a group of analysts tasked to sift through and sort 100 pictures into designated groups. The scale of this proposed situation is not extreme, and if these 100 pictures were placed on a desk for analysts to work through, they would be able to create different groupings, or clusters, in a short amount of time. However, with the sheer amount of data that can be extracted from a battlefield due to the differing sensors that are in play, whether that be drone footage, CCTV imagery, or snapshots from other sources, it would be unusual for only 100 images to be obtained. Instead, take the same situation but increase the number of images from 100 to 10,000. What can be observed is that when the scale of the situation increases, which is likely to occur in an actual operational environment, the difficulty of working through and sorting all 10,000 images increases rapidly.

ISBN: 97819384964-4-8

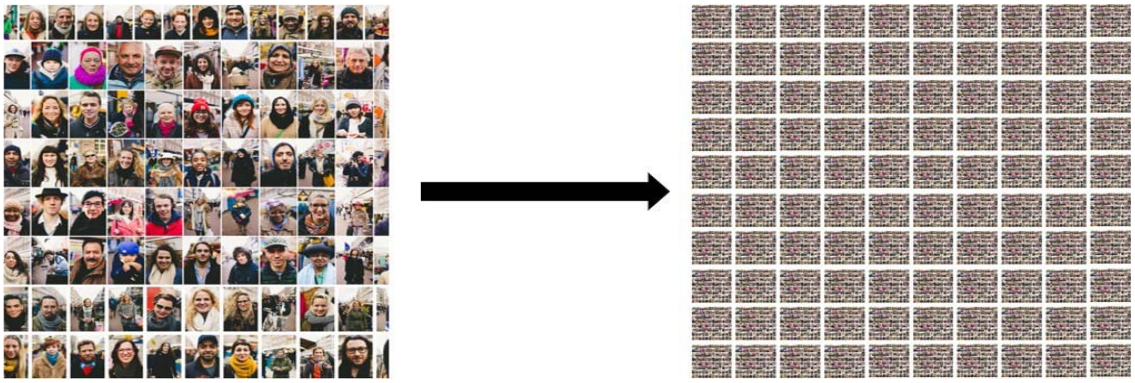

Figure 1. Manually Sorting of Labeling Images Goes from Manageable at 100 Images to Terribly Inefficient at 10,000 Images.

Reference Figure 1 for a visual representation of the difficulty that arises when the number of images increases from 100 to 10,000. It is hard to make out each individual grouping of 100 images, let alone the actual faces within each of those groupings. If a team of analysts was tasked with sifting through all of these images in order to create accurate groupings, it would take this team a long time to sort through all 10,000 images while accurately grouping them into different clusters.

This is where the art of image classification that takes place within something such as target identification can be aided by data science. When an unsupervised classification process such as k-means is paired with experts providing class labels for images tagged to be looked at, it results in a semi-supervised approach to image classification that is more accurate than running an unsupervised approach such as k-means. Additionally, it is more time-efficient than using a strictly user approach where an analyst has to sift through every image, attach a label, and sort these images into their respective groupings.

2. Background

Two approaches for image classification that are thoroughly researched include supervised and unsupervised learning (Kotsiantis, Zaharakis, Pintelas, et al., 2007) (Domadia & Zaveri, 2011). Fully supervised image classification involves training a model on a pre-labeled set of images (Eick, Zeidat, & Zhao, 2004). An issue with this approach is that when images are taken from the wild, there is not a label attached to them. This excludes fully supervised learning from being an option within the proposed situation as without having pre-determined labels, it is unfeasible to implement a fully supervised classification approach.

A different classification method that could be used is an unsupervised learning approach. Unsupervised learning is the opposite of fully supervised learning, where instead of knowing the label of each image, none of the labels are known (Olaode, Naghdy, & Todd, 2014). Unsupervised learning looks at patterns within the data and creates groupings based off of these pat- terns. An unsupervised approach to image classification, k-means, involves extracting data from an image, randomly assigning centroids to an object space representing the data, assigning the data to be classified to the closest centroid using some distance metric such as the Euclidean distance, adjusting the centroids of each group to the center of each cluster, and then repeating this process until the centroids do not move and each data point is not changing clusters (Rouse, 2016). Labels for new images are predicted based off the cluster that this data is closest to, again using a distance metric such as the Euclidean distance. While this could be applied to our problem, the goal of our research is to utilize the knowledge and capabilities of subject matter experts, which is not possible in an unsupervised clustering approach.

Instead, a semi-supervised approach is more fitting for our proposed problem. A semi-supervised approach to image classi- fication involves having a small group of labeled data along with a large amount of unlabeled data to create the clusters (Grira, Crucianu, & Boujemaa, 2004). Unlike fully supervised learning where all of the data labels are known or fully unsupervised learning where none of labels are known, semi-supervised learning falls in the middle of these two typical clustering methods. Seeded k-means clustering (Bair, 2013) is a semi-supervised classification approach similar to an unsupervised approach, but unlike unsupervised k-means clustering, seeded k-means clustering uses labeled data to assign the original location of the cen- troids. We utilize a manipulation of the seeded k-means approach, where labeled data is used to assign the original centroid locations with further labeled data, being classified by a user, allowing for continuous seeding in order to refine the clusters and improve the overall accuracy of the results.

3. Data

3.1. FairFace

The FairFace data set is a conglomeration of faces relating to individuals differing in age, gender, and race (Karkkainen & Joo, 2021) with a total of 500 images. The benefit of this data set is that the faces contained within it are extracted from within the wild. This means that the lighting and angle at which the individual is looking at the camera when paired with the differing ages, genders, and races of the faces within the data are not perfect. This is important as images taken of high valued targets will vary in quality and have inconsistencies within the properties contained in the images, meaning that for the image classification techniques to be useful, our methods need to be able to handle a variety of different properties.

3.2. Celebrity

The Celebrity data set is a face recognition data set taken from Kaggle which contains 31 different classes, each class relating to a different celebrity and each celebrity having roughly 100 images associated with them (Patel, 2019). The point of this data set is for ease of use as well as a way to double check that the clusters which our methodology creates of different faces relate to the correct celebrity.

4. Methodology

We used the FairFace data set to validate a facial vector extraction method on a wide variety of individuals. Images are difficult to work with as the data contained in them consists of a tensor of typically four matrices, with the first three matrices relating to the color value of either red, green, or blue of each pixel within the image and the fourth matrix relating to the transparency of these pixels (HW, 2021). In order to convert the data contained within each image into a usable form, each image was passed through a pre-trained convolution neural network (CNN), specifically MTCNN (Zhu & Ramanan, 2012), to perform facial detection and alignment before being passed through InceptionResnetV1 (Kankam, 2022) pre-trained on vggface2 (Cao, Shen, Xie, Parkhi, & Zisserman, 2018) to extract a face embedding, or a feature vector relating to each face (AIData, 2022). Roughly 500 faces from the FairFace data set were passed through this pipeline with only three images not being able to be processed. These three images consisted of situations where the face was obscured to a point where it was too difficult for the CNNs to recognize that a face was present. The faces from the Celebrity data set were then passed through the same pipeline to extract the facial embeddings relating to these images.

An issue with the facial embeddings extracted from each image was the dimensionality of these vectors as each vector was of length 512. K-means uses some form of distance metric, typically the Euclidean distance between a data point and a centroid location, to assign a data point to a certain cluster. The Euclidean distance for calculating the distance between two points in two dimensions is as follows:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

With the vectors being in 512 dimensions, even though some centroid would technically be closest to a vector being classi- fied, the Euclidean distance could still be extremely far away (Shetty, 2022). In order to combat this, we ran principal component analysis (PCA) (Abdi & Williams, 2010) to reduce the dimensionality down from 512 to 50. With the first 50 principal com- ponents (PCs) being extracted from each facial vector, the first two PCs of 200 images (20 images of 10 different celebrities) were plotted onto a scatter plot represented in Figure 2. Clusters relating to each celebrity appeared to be formed as the points taken from each celebrity were generally grouped together and separate from other celebrities. It is important to note that we were not looking for the best way to conduct image classification. We were concerned with demonstrating the benefit of a semi-supervised approach to image classification in situations where state-of-the-art algorithms may not perform well enough on their own. With this in mind, we utilized only the first two PCs from each facial embedding to conduct further analysis as this made it easy to visualize what was occurring during our semi-supervised approach.

ISBN: 97819384964-4-8

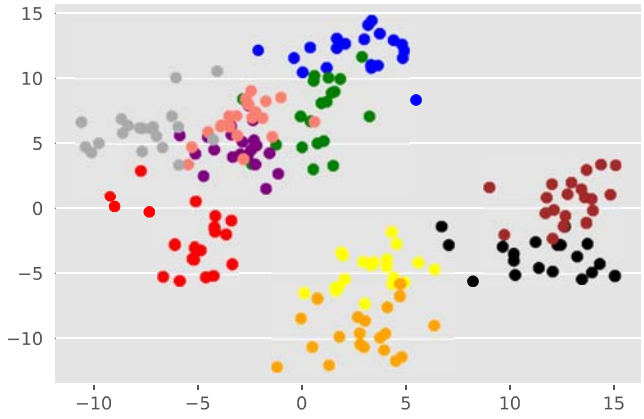

Figure 2. Clusters Formed Within First Two PC's with 10 Different Celebrities

We implemented a k-means algorithm such that we could halt the clustering process to enable a user to insert knowledge into the system. The algorithm prompted users to examine specific images for which there was low confidence in the classification label predicted by k-means. This allowed the user to assign a label to the specified image manually. By allowing user input into the system, our methodology introduced a small subset of labeled images into the process, creating a semi-supervised approach. Twenty different images from five different celebrities resulting in 100 images ranging in age, race, and gender were all selected at random. While we knew the labels of these images, the algorithm did not as only the first two principal components from each of these images was utilized for classification. Another image from each of these celebrities was selected and used as the initial centroid location of each different cluster. Throughout the rest of the process, this image would not shift groups as we had attached a true label to it relating to the corresponding celebrity and cluster it belonged to. We then ran traditional k-means clustering on this data, creating solidified centroid locations for each of the five different clusters.

With the centroid locations solidified, multivariate Gaussian distributions were fit using maximum likelihood estimation (Myung, 2003) relating to each of the five clusters. Each image was given a probability density value associated with each of the five different clusters, and the greatest density value was selected for each image. Within these 100 images, the image with the lowest maximum density value, calculated utilizing Equation 2, was presented to the user along with an image the algorithm knew with certainty belonged to the group the image in question was closest to. Within Equation 2, f corresponds to a multivariate Gaussian pdf, x^* relates to a vector representing the first two PCs of a specific image, and θ_i represents the parameters relating to a specific cluster.

$$min\ [max\ [f(x^*|\theta_i)]] \tag{2}$$

The user was then able to input a 0 if the two presented images did not belong to the same celebrity or a 1 if they did. If a 0 was entered, the next highest density value (i.e., the next most likely class) for the image in question was obtained, and a face belonging to the next most likely class match was presented to the user. The user was again able to input a 0 or a 1, with this process continuing until the user inputted a 1, specifying that the image being classified belonged to the presented class. This process was repeated on another four images, resulting in five images being manually labeled to their respective groups. With these five images possibly changing groups, k-means was run once again with new centroid locations for each group being calculated; however, this time if an image had been locked to a group by a user, it was kept within the user-specified cluster. Even if another centroid was closer in distance than the centroid of the manually labeled cluster, the locked image would remain assigned to the user-predicted group.

With new groups being formed after running k-means a second time, the process of fitting multivariate Gaussian distributions, obtaining predicted density values, and manually assigning prompted images followed by k-means was repeated until every image was manually classified. From this point, the process was repeated another nine times with the same group of celebrities. After 10 iterations were run on one group of five celebrities, a new group of celebrities was selected and the process was repeated another 10 times. This was done until 50 total iterations were completed.

5. Results

Referencing Figure 3, the accuracy scores of the strictly user approach is dictated by the bold black line, the mean of the unsupervised approach is dictated by the bold blue line, and the mean of the semi-supervised approach is dictated by the bold red line. It is important to note that the x-axis, images observed, dictates the number of images that have been manually labeled by the user. As previously stated, an unsupervised approach does not have any user input as there are no labels manually assigned. Instead, the algorithm simply creates clusters based off of what it dictates the most accurate groupings to be. This is why the blue lines, indicating an unsupervised approach, run horizontally along the plot. For the user-only approach, each image was manually labeled without any original grouping happening beforehand.

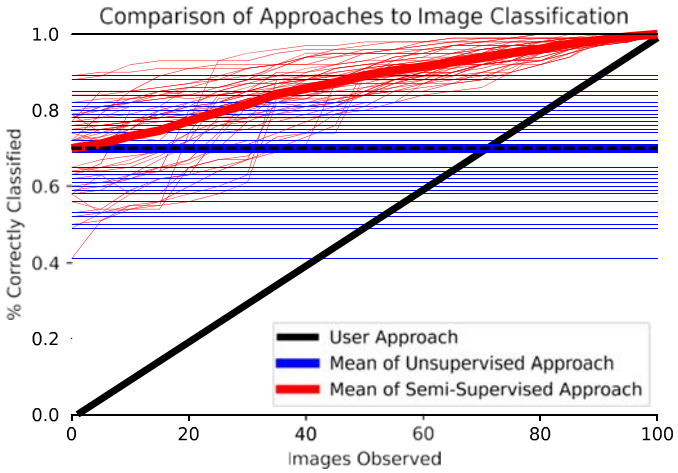

Figure 3. Results of User, Unsupervised, and Semi-Supervised Approaches

Figure 3 shows that the unsupervised approach has an average accuracy of around 70%. The user approach starts at 0% and arrives at 100% after all the images have been manually labeled. Additionally, the semi-supervised approach reaches an accuracy of roughly 90% when around half of the images are manually classified. The dashed black, representing the average increase in the percentage of correctly classified images that would result if a user manually labeled every image in a random fashion after unsupervised clustering had taken place, demonstrates that even if you start the user off with a good initial clustering, the semi-supervised approach still does better faster by interlocking the human and machine efforts. During each semi-supervised iteration, there is a point where the percentage of correctly classified images spikes as seen within the light red lines. The mean of the semi-supervised approach does not show this spike as it has been averaged out. On average, a jump of 5% or greater occurred after the fifth group of images was observed, or 25 total images being manually labeled by the user. Throughout the 50 total iterations, a jump of 5% or greater occurred a total of 34 times, with the majority of these jumps occurring at the fourth, fifth, sixth, or seventh set of images observed.

6. Discussion

6.1. Significance

The key takeaway from our results is that the semi-supervised approach outperformed both the strictly unsupervised approach and user-only approach. The semi-supervised approach will never perform worse than the unsupervised approach in the long run under the assumption that the user manually labeling the images is never wrong. This is demonstrated in Figure 3 where the percent of correctly classified images for an individual semi-supervised approach dips below the unsupervised approach when only a couple of images have been manually labeled, but this will always correct itself as more images are manually labeled. While the user-only approach will eventually reach an accuracy of 100%, it takes a long time for this to occur. The semi-supervised approach was able to reach an accuracy of around 90% on average by looking at roughly half of the images compared to a user having to manually label 90% of the images in the user-only approach, enabling the user to obtain a much higher accuracy than the unsupervised approach in far less time than the user-only approach.

This demonstrates how the semi-supervised approach to image classification operates between the two extremes of

strictly unsupervised classification and strictly user classification. Functioning in a middle ground, the semi-supervised approach per- forms better than the unsupervised approach and saves valuable time for a subject matter expert that is lost in the strictly user approach, fusing the pros of these two methods together while limiting the cons.

6.2. Limitations

We recognize that our method of image classification using a semi-supervised approach is not the most efficient way to conduct image classification as we needed our original results to be poor to show improvement.

6.2.1. Overlap

As the number of possible classes grows, the overlap between the clusters also rises. We used five total classes for each classification iteration and when there were individuals being classified who looked similar, the original clusters that were formed by the unsupervised classification approach had a very low accuracy score. The overlap between classes limited the benefit of running an unsupervised k-means classification algorithm beforehand as a greater number of images needed to be manually classified by a user to achieve a high accuracy score.

6.2.2. User Error

We worked under the assumption that an individual would never classify an image incorrectly. If this was to occur, the clusters being formed would shift in the wrong direction. Our methodology prevents an image from being classified twice by the user, leaving no possibility for a mistake to be corrected.

6.2.3. Outliers

Our current practice of selecting the image with the lowest maximum density value results in the outliers being selected to be manually labeled first by the user. This may not be the best practice as there could be other images which may add more information to the process, such as "horizon" images which are located on the edges of two clusters and have similar maximum density values.

7. Conclusion

Our research shows how a semi-supervised approach to image classification can perform with a higher accuracy than an unsupervised approach while being quicker than having a user manually label every image. While our methodology did not utilize a state of the art image classifier, we used k-means with two PCs to provide an illustrative example of how certain practices performed by an individual can be improved upon using data science techniques. The next step would be to use more powerful techniques involving image classification to show how these techniques, when paired with user knowledge, can result in an improved approach.

8. References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433–459.

AIData. (2022). What is facial recognition? Applications and how it works. Retrieved from https://www.telusinternational.com /insights/ai-data/article/what-is-facial-recognition.

Army, U. (2020). 35a: Military intelligence officer. Retrieved from https://www.goarmy.com/careers-and-jobs/career-match/signal-intelligence/languages-code/35a-military-intelliigence-officer.html. United States Army.

Bair, E. (2013). Semi-supervised clustering methods. Wiley Interdisciplinary Reviews: *Computational Statistics, 5*(5), 349–361.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)* (pp. 67–74).

Domadia, S. G., & Zaveri, T. (2011). Comparative analysis of unsupervised and supervised image classification techniques. In *Proceeding of national conference on recent trends in engineering & technology* (pp. 1–5).

Eick, C. F., Zeidat, N., & Zhao, Z. (2004). Supervised clustering-algorithms and benefits. In *16th ieee international conference on tools with artificial intelligence* (pp. 774–776).

Grira, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey. *A review of*

machine learning techniques for processing multimedia content, 1(2004), 9–16.

HW, B. (2021). What the heck does it mean to make an image a tensor? Retrieved from https://dev.to/bekahhw/what-the-heck-does-it-mean-to-make-an-image-a-tensor-4feh:

Kankam, N. (2022). *Understanding inception-resnet v1 architecture*. Retrieved from https://iq.opengenus.org/inception-resnet-v1/.

Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1548–1558).

Kotsiantis, S. B., Zaharakis, I., Pintelas, P., et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*(1), 3–24.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. Journal of mathematical *Psychology, 47*(1), 90–100.

Olaode, A., Naghdy, G., & Todd, C. (2014). Unsupervised classification of images: a review. *International Journal of Image Processing, 8*(5), 325–342.

Patel, V. (2019). *Face recognition dataset: Kaggle*. Retrieved from https://www.kaggle.com/datasets/vasukipatel/face-recognition-dataset

Rouse, M. (2016). K-means clustering. Retrieved from https://www.techopedia.com/definition/32057/k-means-clustering.

Shetty, B. (2022). *What is the curse of dimensionality?* Retrieved from https://builtin.com/data-science/curse-dimensionality.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *2012 ieee conference on computer vision and pattern recognition* (pp. 2879–2886).

ISBN: 97819384964-4-8