Type of the Paper (Article, Review, Communication, etc.)

Improvement for Large-Scale Image Data using Fuzzy Rough C-Mean Based Unsupervised CNN Clustering: Case Study designbyhumans.com

Tuan Anh Tran 1, Quy Ban Tran 1,2 and Firstname Lastname 2,\*

|  |
| --- |
| **Citation:** Lastname, F.; Lastname, F.; Lastname, F. Title. *Entropy* **2022**, *24*, x. https://doi.org/10.3390/xxxxx  Academic Editor: Firstname Lastname  Received: date  Accepted: date  Published: date  **Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.    **Copyright:** © 2022 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |

1 Affiliation 1; e-mail@e-mail.com

2 Affiliation 2; e-mail@e-mail.com

**\*** Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials)

**Abstract:** Clustering analysis, specifically, for large image data are increasingly being applied in various fields such as finance, risk management, prediction, etc., and has been an interesting subject in many science discussions. Deep learning, a widely used approach, along with classical methods is being used to address sophisticated classification problems stem from real world cases. In this study we will be taking various approaches to the classification problems and also, measure their effectiveness by combining different methods using the results taken from different scenarios in this study. Initial evaluations will be conducted on a dataset that we have been able to collect from the website designbyhumans.com and the result will be analyzed in details with each technique and will be compared with other similar and different approaches.

**Keywords:** FRCM; CNN; keyword 3 (List three to ten pertinent keywords specific to the article yet reasonably common within the subject discipline.)

1. Introduction

Since the beginning of the information technology era, image processing has become an important research field, its necessity has been heavily implied in computer vision applications, where even a single image can potentially contain valuable information. 3.2 billion images are uploaded to the Internet every day with different purposes, that huge number indicated the demand of integrating large scale image data into a number of different fields. A huge number of efforts have been made to meet those artificial demands, such as reducing a number of samples from the already enlarged dataset or encoding image features within the dataset as there might and usually exists redundant and noise samples scattered throughout the dataset with varied magnitude.

Unsupervised learning is a popular machine learning tool to visualize the structure beneath the huge dataset. An unsupervised method must be able to reproduce the result without too much prior knowledge from previous learning from unlabeled datasets, which is commonplace in the real world and typically unknown before applying cluster analysis techniques. Distinctive or so-called “hard” clustering methods such as hierarchical, density-based, centroid-based and graph theoretical have been used extensively for prediction on these structures with commendable precision. However, with the existence of high quality image data with great depth to their perception, these “hard” methods are left exposed to a huge challenge: The amount of immeasurable vagueness, uncertainty or overlapping of samples from the clusters is nearly impossible to make precise predictions to each sample.

Because of such uncertainties, Rough Set and Fuzzy Set theory have been introduced and implemented as cluster analysis methods, so-called “soft” methods opposed to “hard” methods in order to highlight the blurry nature of the data that might be an oversight to well-known “hard” methods. Two prominent methods are Fuzzy C-means and Rough C-means, which use probability-based weighting values to identify the clusters. Rough C-means describes a cluster with a centroid which is initialized by choosing data points randomly, and a pair of lower and upper approximation with different weighting values. Fuzzy C-means on the other hand use weight calculation formula to impose different weighting values based on distances to cluster centers. Fuzzy Rough C-means (FRCM) is the combination of these abovementioned to overcome both the overlapping nature and uncertainty of image data. It incorporated elements of the two: A center, a crisp lower approximation and a fuzzy boundary approximation, classifying based on rough approximation and calculating based on fuzzy approximation.

Deep learning is also a trending topic in recent years because of its specialization in solving certain problems, especially in analyzing large image data. Some architectures that represent its robustness include Deep Belief Network, Deep Boltzmann Machine, Deep Autoencoder and Convolutional Neural Networks. For video and imagery applications, we’ve already had CNN-based architectures: AlexNet, ResNet, VGG and FCN. These architectures demonstrated robust accuracy in image data classification tasks with sufficient labeled training set.

Clustering algorithm’s performance depends greatly on noise reduction and feature representation power. Deep learning architectures rely on accurate labeled data, which is usually not available in the real world applications and does not applicable in cluster analysis applications. Therefore, the model can be pre-trained on existing dataset with reliable labeling process and using transfer learning and proper tuning to be able to identify such inconsistency.

Many efforts have been done to combine cluster analysis with representation learning. Hsu[citation] proposed CNN-based joint clustering in which Mini-batch K-Means algorithm is executed to assign cluster labels. However, it carries the cons of the hard-clustering method, is the deterministic model that incomprehensible with the inconsistency of the image dataset. Xie[citation] proposed Autoencoder-based deep learning, followed by K-Means to get initial clusters, but the autoencoder cannot properly learn representative features for high dimensional data.

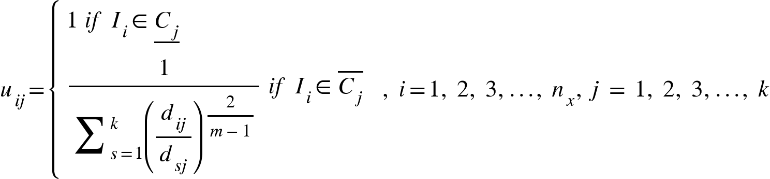
In this paper, we will be using combined CNN and FRCM architecture provided from the paper[citation] but with some modification in order to fit in the basic concepts that has been cited in the given paper using our own dataset from the website designbyhumans.com and we will be comparing the performance between the dataset and MNIST dataset.

* 1. Background
     1. *Fuzzy Rough C-means (FRCM) clustering algorithm:*

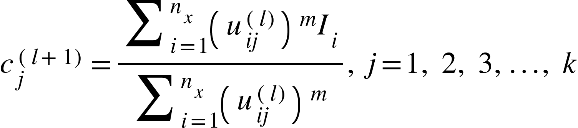
As stated by Hu et al. [citation], the FRCM algorithm is a combination of both Fuzzyand Rough C-means. FCM maps each data points into a membership matrix which ranges from 0 to 1, each points belong to some or all the clusters to some degree probabilistically and calculate the centroids based on distances to each of the cluster centers that have been initialized by random sampling. RCM classifies the points into two parts: The lower approximation, upper approximation; those who belong in the lower approximation are guaranteed to be a data point of that cluster, and the ones in the upper part belong to a cluster by some extent with respect to their different weighting values. Inspired by these concepts, FRCM integrated all of these elements and imposed fuzzy membership values of each sample to the lower and upper area of the clusters.

Let a set of image data I space equals space open curly brackets I subscript 1 comma space I subscript 2 comma space I subscript 3 comma... comma I subscript n subscript x end subscript close curly brackets element of R to the power of d , where d is the dimension of the data points. Each cluster C subscript j open parentheses j equals 1 comma space 2 comma space 3 comma... comma space n close parentheses, where k denotes the number of clusters, is regarded as a rough set. The data points are categorized into the lower approximation bottom enclose C subscript j end enclose and the upper approximation top enclose C subscript j end enclose . Let c space equals space open curly brackets c subscript 1 comma space c subscript 2 comma space c subscript 3 comma... comma space c subscript k close curly brackets be a vector composed of k centers of clusters, where c subscript j element of R to the power of d. The points in lower approximation are guaranteed to be in the clusters and take the same weight value, while the ones that are in the upper region have diverse effects on the centers and clusters, therefore different weighting values must be imposed on only the points belong to the upper region to compute new cluster centers.

Let u equals open curly brackets u subscript i open parentheses j close parentheses close curly brackets subscript n subscript x x k end subscript be a membership matrix, we have the membership function:

 (1)

The exponent m > 1 is the parameter to change the weighting impact of membership values, usually in range from 1.5 to 2.5 as different methods suggested. The new cluster center is computed according to the function:

 (2)

And the objective function for FRCM as follows:

J subscript m superscript open parentheses l close parentheses end superscript open parentheses u comma space c close parentheses equals sum from i equals 1 to n subscript x of sum from j equals 1 to k of open parentheses u subscript i j end subscript close parentheses to the power of m open double vertical bar I subscript i minus c subscript j close double vertical bar squared (3)

To wrap up, the algorithm is formulated as follows:

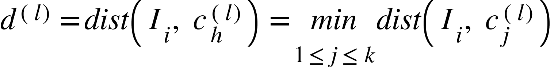
**Input:** Unlabeled data I, number of clusters k, threshold parameter T, exponent index m, stop criterion epsilon.

**Output:** Membership matrix u, k cluster centers

**Step 0:** Let l equals 0, initialize membership matrix u randomly

**Step 1:** Calculate pairwise distances between each data points and calculate cluster centers c subscript j superscript open parentheses l close parentheses end superscript equals 1 comma space 2 comma space 3 comma... comma space k using randomly initialized membership matrix u and equation (2)

**Step 2:** Assign data points to the approximations:

1. Calculate its closest center c subscript h superscript open parentheses l close parentheses end superscriptand A to the power of open parentheses l close parentheses end exponentas follows:
   1.  (4)
   2. A equals open curly brackets for all s comma space s equals 1 comma space 2 comma space 3 comma... comma space k comma space s not equal to h colon space d subscript i s end subscript superscript open parentheses l close parentheses end superscript minus d subscript i h end subscript superscript open parentheses l close parentheses end superscript less or equal than T close curly brackets
2. Classify data points based on A:
   1. I f space A not equal to empty set comma space t h e n space I subscript i element of top enclose C subscript h end enclose space comma space I subscript i element of bottom enclose C subscript s end enclose space a n d space I subscript i not an element of bottom enclose C subscript l end enclose comma space l equals 1 comma space 2 comma space 3 comma... comma space k
   2. I f space A equals empty set comma space t h e n space I subscript i element of bottom enclose C subscript h end enclose comma space I subscript i element of top enclose C subscript h end enclose

**Step 3:** Compute new membership values using equation (1)

**Step 4:** Compute cost using equation (3). If the cost meets the stopping criterion, stop the algorithm. Otherwise repeat step 2.

*1.1.2. Fuzzy Rough UCNN Clustering Architecture:*

To be able to enhance clustering algorithm’s performance, unsupervised CNN is integrated along with Fuzzy Rough C-Means clustering algorithm. There are two parts in this architecture: The pre-clustering stage and further joint clustering and representation learning. The clusters are updated by using the FRCM algorithm in the forward pass of the architecture and optimized by the stochastic gradient descent in the backward pass.

In the pre-clustering stage, the size of the dataset will determine how large the respective multi-convolutional layers will be and will generally requires large-scale networks in case the dataset is huge enough. For the imagenet dataset used in this paper, the AlexNet architecture is proven to be effective enough to implement the training. The architecture consists of 5 convolutional layers (Conv1 – Conv5) used in AlexNet, followed along by 3 adjustment layers (Conv6, Conv7, CConv) with channel number 6144, 2048 and k, connected with a fully connected (FC) layer and a softmax layer to extract the image features and predict the cluster labels.

[Image of the architecture goes here]

1.2. Related Work

Hu et al. and Yu et al.[citation] proposed a clustering algorithm which is the combination of Fuzzy C-means and Rough C-means that perceive real world applications of unstable and blurry data. Each data point is categorized based on threshold value and its membership value into lower and upper approximation. The centers are re-calculated using randomly created membership matrix and new membership value, correspond with data point’s coordinates and membership value, along with an exponent index value, usually stable at 2.0.

Hsu et al. and Lin et al.[citation] proposed a clustering CNN architecture and representation learning method to cluster image data based on their extracted visual representations. A clustering algorithm is utilized to support the CCNN by assigning the labels to each image in the dataset as truth labels and cluster the extracted salient features at each selected output. The CCNN is the primary method as it uses randomly selected samples to feed forward and updates the extracted features of those samples and fine tune its parameters.

Riaz et al., Arshad et al., Jiao et al.[citation] utilized a CCNN architecture based upon FRCM for large image data problem. The architecture has been proven to be better than existing methods for image data clustering problem

2. Materials and Methods

*2.1. Image data pre-processing*

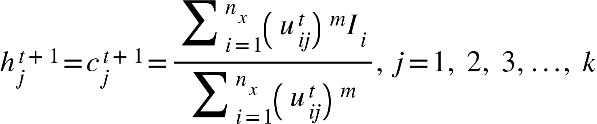
To increase sample variety during the pre-clustering process, the data is pre-processed by using random flip horizontally and random crop at 227 by 227 pixels, as standards before deploying AlexNet to train Conv1 – Conv5 parameters.

*2.2. Cluster centroids calculating*

Let I equals open curly brackets I subscript 1 comma space I subscript 2 comma space I subscript 3 comma... comma I subscript n subscript x end subscript close curly brackets be the set of n subscript x images. We will be using the membership matrix from the result of the ground truth from applying FRCM to the dataset with the extracted centroid features using UCNN H subscript j superscript open parentheses t close parentheses end superscript equals open curly brackets h subscript 1 superscript t comma space h subscript 2 superscript t comma space h subscript 3 superscript t comma... comma space h subscript k superscript t close curly brackets element of H subscript F C end subscript from the FC layer as extracted dataset features initial centroids in order to minimize the gap between using independent features and dataset. The FRCM is performed to update the cluster centroids by the objective function:

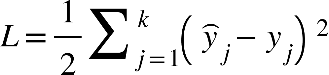
J subscript m open parentheses u comma space h close parentheses equals sum from i equals 1 to n subscript x of sum from j equals 1 to k of open parentheses u subscript i j end subscript close parentheses to the power of m open double vertical bar I subscript i minus h subscript j close double vertical bar squared

And the centroids are calculated by the function:



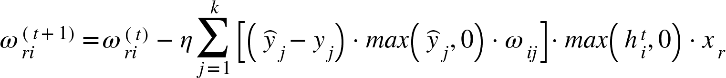
*2.3. Representation learning*

By using the UCNN architecture, features of the images in the dataset are extracted by using the FC layer and salient features are extracted using CConv layer. To learn the parameter theta open parentheses omega subscript r i end subscript comma space omega subscript i j end subscript close parenthesesof FC and softmax layer, we will be using SGD process, where omega subscript r i end subscript belongs to the FC layer and omega subscript i j end subscript belongs to the softmax layer, respectively. We will start with the objective function:



where k is denoted as the number of clusters, y with overparenthesis on top subscript j is the predicted jth cluster label using the UCNN and y subscript j is the predicted jth cluster label using FRCM as pseudo ground truth. Then by using chain rule and calculating gradient, we have the functions for updating FC and softmax layers:

omega subscript i j end subscript superscript open parentheses t plus 1 close parentheses end superscript equals omega subscript i j end subscript superscript open parentheses t close parentheses end superscript minus eta open parentheses y with overparenthesis on top subscript j minus y subscript j close parentheses times m a x open parentheses y with overparenthesis on top subscript j comma 0 close parentheses times h subscript i superscript open parentheses t close parentheses end superscript (5)

 (6)

*2.4. The complete algorithm:*

Input:

Image dataset I, k number of clusters, randomly selected data

Learning rate

Max iteration

Extracted image features of centroids of FC layer.

Output:

Final cluster centroids

Final weights ()

1. For t = 1 to do:
2. Calculate cluster label using FRCM as ground truth on randomly utilized centroids to speedup algorithm’s convergence
3. Forward feed the centroids and extract the features from UCNN’s clustering layer and FC layer.
4. Update cluster centroids by using FRCM on randomly sampled centroids in the FC layer and find the predicted cluster label based on the updated centroids.
5. Update the weights of FC layer and softmax layer by using function (5) and (6)
6. Fine tune the UCNN by using the objective function

*2.5. Experiment:*

*2.5.1. Data preparations:*

In this paper, Python 3.9 was utilized as our programming tool because of its compatibility with various libraries and frameworks. ILSVRC19 train set in ImageNet is used in pre-training stage with our representation of AlexNet; it consists of over 1.2 million training images of size 256 x 256-pixels collected from 1000 categories. As mentioned above, our provided by DesignedByHumans dataset is used to evaluate the performance of this approach, which consists of over 52000 images of size 1200 x 1200-pixels downscaled to 28 x 28-pixels collected from 69 different categories includes sub-categories and possibly belong to many categories as once, which emulates real world example of unstable data for semi-supervised learning method.

The hardware we will be using is a personal computer using Ryzen 7 4800HS CPU and NVIDIA GTX 1650 with 16GB of RAM updated with newest drivers on Tensorflow 2.9.

**Table 1.** Description of Datasets

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **No. of Samples** | **No. of Classes** | **Image size** |
| ILSVRC19 | 1,281,167 | 1000 | 256 x 256 |
| DesignByHumans dataset | 52,361 | 69 | 28 x 28 |

*2.5.2. Evaluation metrics:*

In order to evaluate the efficiency of the method, we will be using metrics that are often used to evaluate clustering algorithms: Normal Mutual Information (NMI), Clustering accuracy, mean F-Measure score and Area Under the Curve score (AUC).

NMI is defined as:

Where C is the truth label, Y is the predicted label, H(x) stands for Entropy and I(C,Y) = H(C) – H(C/Y) indicates the mutual information between C and Y. This is also the metric we used to optimize the number of clusters in the dataset

Clustering accuracy is used to evaluate the performance in each class independently

m is the number of assumed classes

The mean F-Measure score is defined for multi-class classifications as:

Where F-Measure is calculated using:

Mean AUC is the average value of pairwise AUC values of all pairs of classes:

*2.5.3. Comparison and evaluate:*

To evaluate the performance of the method and compare it with other methods using the same dataset, we will be conducting our tests in comparison with KMeans since 1. KMeans is the most popular clustering algorithm using in most of the comparison and 2. It’s performance in synthetic and large data is still questionable in terms of overall performance

*2.5.4. Implementations and practices:*

We began using the pre-trained AlexNet on the updated ILSVRC19 training set of ImageNet as our basic convolutional model and data augmentation methods: random flip left and right as well as random cropping samples to size 227x227 during the pre-training process to increase samples variety.

The proposed model consists of 5 convolutional layers taken from the pre-trained AlexNet, with the same configurations as the original model, concatenated with 2 adjustment layers with filter size 6144 & 2048, with kernel size 3 x 3 and a clustering convolutional layer with k channel size and the same kernel size as the adjustment layers, and followed by a global max pooling layer with 1 x k output, k is denoted as the number of clusters.

We evaluated the model on our dataset, consists of over 50000 RGB images of size 1200 x 1200 . The output of the FC layer is considered the centroids of the clusters, and is updated in the forward propagation stage using FRCM and SGD in the backward propagation stage.

Our dataset is resized to 28 x 28 due to GPU memory constraints, as well as the model.

3. Results

*3. 1. Performance measures:*

*3.2. Computational time:*

*3.3. Impact on performance:*

4. Discussion

Dataset quality is our main concern because as stated, the number of features in a HD quality image is demanding for the hardware, and the fact that we had to face constraints in computational power reflected this. The model is also taxing on the GPU memory that the performance may suffered heavily.

FRCM is also an inconsistent method, as it tends not to converge in many situations during our implementation. Although the Fuzzy C-means is guaranteed to reach a convergence point, there is no hypothetical guarantee that FRCM algorithm reach convergence for large and inconsistent dataset as ours

5. Conclusions

In this documentation, we have applied the CNN model on our synthetic dataset based on the Fuzzy Rough C-means clustering algorithm. The algorithm can provide and be improved for better results on large scale, high dimensions image data by using iterations between updating cluster centroids using FRCM algorithm and fine-tuning process of the initial model. The CNN model was able to effectively extract salient features from the clustering layer as cluster centroids and features and updated in the forward pass. Comparison with KMeans proved the effectiveness of the method on selective dataset and the robustness of the algorithm with one of the most popular unsupervised learning methods. However, its uncertainty and defect in the operation will be obstacles that need to be address and improved upon. Being self-adaptive will be the objective of this learning method in the future

6. Patents

This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript.

**Supplementary Materials:** The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.” Please turn to the [CRediT taxonomy](https://img.mdpi.org/data/contributor-role-instruction.pdf) for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER, grant number XXX” and “The APC was funded by XXX”. Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding. Any errors may affect your future funding.

**Data Availability Statement:** In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section “MDPI Research Data Policies” at https://www.mdpi.com/ethics. If the study did not report any data, you might add “Not applicable” here.

**Acknowledgments:** In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

**Conflicts of Interest:** Declare conflicts of interest or state “The authors declare no conflict of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state “The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results”.

**Appendix A**

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

**Appendix B**

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled starting with “A”—e.g., Figure A1, Figure A2, etc.

References

References must be numbered in order of appearance in the text (including citations in tables and legends) and listed individually at the end of the manuscript. We recommend preparing the references with a bibliography software package, such as EndNote, ReferenceManager or Zotero to avoid typing mistakes and duplicated references. Include the digital object identifier (DOI) for all references where available.

Citations and references in the Supplementary Materials are permitted provided that they also appear in the reference list here.

In the text, reference numbers should be placed in square brackets [ ] and placed before the punctuation; for example [1], [1–3] or [1,3]. For embedded citations in the text with pagination, use both parentheses and brackets to indicate the reference number and page numbers; for example [5] (p. 10), or [6] (pp. 101–105).

1. Author 1, A.B.; Author 2, C.D. Title of the article. *Abbreviated Journal Name* **Year**, *Volume*, page range.
2. Author 1, A.; Author 2, B. Title of the chapter. In *Book Title*, 2nd ed.; Editor 1, A., Editor 2, B., Eds.; Publisher: Publisher Location, Country, 2007; Volume 3, pp. 154–196.
3. Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, 2008; pp. 154–196.
4. Author 1, A.B.; Author 2, C. Title of Unpublished Work. *Abbreviated Journal Name* year, *phrase indicating stage of publication (submitted; accepted; in press)*.
5. Author 1, A.B. (University, City, State, Country); Author 2, C. (Institute, City, State, Country). Personal communication, 2012.
6. Author 1, A.B.; Author 2, C.D.; Author 3, E.F. Title of Presentation. In Proceedings of the Name of the Conference, Location of Conference, Country, Date of Conference (Day Month Year).
7. Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.
8. Title of Site. Available online: URL (accessed on Day Month Year).