**PROPOSAL / THUYẾT MINH ĐỀ CƯƠNG**

**FOR RESEARCH PROJECT**

|  |  |
| --- | --- |
| Project name (Vietnamese) | Phương pháp K-means ứng dụng phân cụm dữ liệu ảnh lớn, thí nghiệm với tập dữ liệu áo |
| Project name (English) | Toward K-means approach to cluster the big image data, evidence in shirt dataset |
| Sub-committee |  |
| Group name | FU-Lab: Data Mining - Big Data |

1. **Introduction / Giới thiệu tổng quát (Abstract – 300 words)**

*Project summary: problems, objectives, research methodology (maximum 300 words) / Giới thiệu tóm tắt về đề tài (nêu vấn đề, mục tiêu và phương pháp nghiên cứu, không quá 300 từ)*

**Problems:**

* There has been a rapid increase in digital images volume in recent years, so it takes a lot of effort to understand dataset characteristics or annotate labels manually.
* **Time:** It is time-consuming to view all images to understand the nature of the dataset as the size can be huge
* **Cost:** Annotating labels by hand can require a sizeable amount of money

**Objectives:**

Apply Kmeans method to an extensive image dataset to group similar images to a cluster to:

* + Reduce time and cost to grasp an understanding of the nature of the dataset
  + Be able to use the result in other image processing tasks

Make a comprehensive survey of these methods on speed and accuracy to find the best approach.

Make a Web-based application to visual clustering results.

**Methodology:**

1. **Overview of Research situation and the necessity of the project / Tổng quan tình hình nghiên cứu và sự cần thiết tiến hành nghiên cứu**
   1. 2.1.Literature review / Tình hình nghiên cứu trong và ngoài nước

*Review and discuss related researches by Vietnamese and foreign scientists. / Nêu rõ những nghiên cứu do các nhà khoa học nước ngoài, trong nước tiến hành theo hướng nghiên cứu của đề tài và các kết quả đạt được, đồng thời nhận xét về những kết quả đã nêu.*

*Review the applications of related works in the reality/ Nêu rõ ứng dụng của các nghiên cứu này trên thực tế.*

* 1. **Clustering:** In general, the objective of clustering is to divide data into groups of similar objects, and clustering can be classified into four categories: partitioning, hierarchical, density-based, and grid-based methods [1] The authors in [2] evaluated clustering methods in image segmentation and proposed that in an extensive dataset, partitioning methods are preferred as due to their high computing efficiency and low time complexity and partitioning methods are performed well in their experiments

**Feature descriptor:** A feature descriptor is an algorithm that takes an image and outputs feature descriptors/feature vectors, encoding interesting information into a series of numbers that can differentiate one feature from another. There are global and local feature descriptors. Each image uses a global feature descriptor and gets one descriptor, while each image gets many descriptors using a local feature descriptor.

* 1. **Global and local feature descriptor:**
  2. There are several local feature descriptor: Gradient-based methods such as SIFT or SURF and their variants, intensity-based methods like ORB [4] In a comparison of handcrafted features, SIFT usually get the best result and ORB the fastest ways [5,6]
  3. Also introduced in [4], deep global feature learning is a method based on supervised learning using neural networks. The most common global feature learning model is Convolutional Neural Network (CNN or Convnet). Its well-scalability to high dimensional data led to a breakthrough in computer vision and object recognition [4]. Many CNN-based frameworks have been built specifically for image processing tasks, such as [12]. An autoencoder based on CNN has all of the cons. It can be multi-purposely used within deep clustering and data mining tasks, in which both preprocess stages involve cleaning and denoising high dimensional data.
  4. **Feature reduction:** Feature reduction reduces the number of features in a resource-heavy computation without losing important information. There are two types of feature reduction: feature selection and feature extraction. The former reduces by selecting a subset of features without transforming, while the second transforms the original features to create other elements that should be more significant. [9].
  5. **Bag of Visual Word:** The general idea of a bag of visual words (BOVW) is to represent an image as a set of features consisting of key points and descriptors. We construct vocabularies and represent each image as a frequency histogram of features that are in the image. From the frequency histogram, later, we can find similar images or predict the category of the image [3]

**Deep clustering model:**

Deep clustering model consists of three stages: The first stage involves deep feature learning via a deep neural network, the second stage is to reduce the lower feature representations to visualizable data, and finally, use unsupervised learning to evaluate the dimensional-reduced dataset based on their low dimensional features.[10]

**Clustering method usage in deep clustering models: DBSCAN vs KMeans**

Most deep clustering models combine both supervised and unsupervised learning in order to visualize and evaluate the data, therefore provide general knowledge and insights about the related data. CNN models are considered state of the art and have seen common use in most deep clustering models as deep feature learning method.[4]

However, when it comes to clustering algorithms, 2 methods are always come to mind: density-based and partitioning, and one question remains: Which one of these methods is best used within a deep clustering model ?

In order to answer this question, we have conducted an experiment between two representatives of these above-mentioned methods: DBSCAN and KMeans, explaining our choices in choosing clustering model for our deep learning model:

1. Training and compare the accuracy between the ground truth and predicted labels using Fashion MNIST dataset
2. The same as above, but with our custom dataset

The dataset:

We decided to use 1 common dataset and 1 of our own dataset to provide a general comparison between the methods, since Fashion MNIST is very commonplace in most research papers, and our own dataset to confirm the speculations of how each methods perform such as the implementation in [11]:

1. Fashion MNIST, with 10 categories containing 60000 images of size 28x28
2. Our custom dataset, a collection of shirt designs comprises of 55207 images of size 1200x1200, downscaled to 28x28 to fit the same deep learning model

The implementations:

The number of cluster centers we chose for KMeans algorithm was 7, estimated using elbow method. For DBSCAN, the optimal value for eps is 0.01 using our eps estimator plot. The number of minimum neighbors is always 4

The results:

1. **With Fashion MNIST:**

|  |  |
| --- | --- |
| DBSCAN | KMeans |
|  |  |

On this prepared dataset, KMeans predicted 7 clusters with the accuracy of 57.21%, whilst DBSCAN predicted 4 clusters with the accuracy of 21.01%

1. **With our custom dataset:**

|  |  |
| --- | --- |
| **DBSCAN** | **KMeans** |
|  |  |

After conducting the experiment on our custom dataset, we decided to rely on accuracy result on Fashion MNIST for our final verdict since the ground truth labels in our custom dataset are unreliable due to a large number of outlier labels, and an image may and can be fitted in multiple categories. This result should be taken as the actual result conducted on a real world dataset to give an insights onto how each of these methods actually identifies clusters.

We also acknowledged that other deep clustering models such as [11] have achieved better results on the prepared dataset, but due to outer circumstances, we weren’t able to finish our own clustering method to make a better comparison. We resorted to KMeans due to its flexibility to different datasets

* 1. 2.2. The necessity of the project / Sự cần thiết tiến hành nghiên cứu

*Clearly define the problem or hypothesis to be addressed. / Nêu rõ vấn đề mà đề tài tập trung giải quyết.*

*Originality, relevance and scientific significance of the question under investigation. / Phân tích tính mới, tính thời sự, ý nghĩa khoa học và sự cần thiết của vấn đề cần nghiên cứu.*

There are papers about clustering in some image processing tasks like image segmentation [1], comparing different feature descriptors [6,7], meaning many methods are available to cluster a big image dataset. But in reality, there are only papers about single cluster methods for big image datasets [11]. Therefore, there is a need for a comprehensive comparison of these clustering models.

We implement two K-means based clustering models on a 31-GB shirt dataset mined from designbyhumans website[13] and compare their speed and accuracy.

The first clustering models using Bag of Visual Words to cluster images can be divided into several steps:

1, Using a local descriptor to get a list of all descriptors of all images.

2, Applying feature reduction technique to reduce the dimensions of the list[8]

3, Clustering all the descriptors, then calculating each cluster's representation descriptor to form a Bag of Visual Words.

4, Changing images to vectors based on the created Bag of Visual Words.

5, Clustering images.

The second clustering model using the Deep Learning method is based on the following steps:

1. Train a multi-purpose convolutional autoencoder for feature learning and image denoise. Then use an encoder consisting of 3 convolutional layers, an embedded layer, and a decoder to reconstruct the data.[11]
2. Transform the data to lower feature representations using the encoder. Then apply t-SNE to reduce further the data favoring clustering algorithms.[11]
3. Using the KMeans algorithm to cluster the dimensionality-reduced data and evaluate the visual representation of the low dimensional data
   1. 2.3. Probability of success / Khả năng thành công

*Analyze the probability of success of the project (solution found or improvement of problems/ questions under investigation). / Phân tích khả năng thành công của đề tài (giải quyết hoặc cải thiện được vấn đề nghiên cứu).*

The project has a high probability of success because we have studied technology very carefully, mastered the technology, we have also successfully built models, designed algorithms, and experimented with these algorithms.

1. **Objectives / Mục tiêu của đề tài**

*Clearly outline the topic's objectives as basis for determining research content and implementation plan. / Nêu rõ mục tiêu cần đạt được của đề tài, làm cơ sở xác định nội dung nghiên cứu và kế hoạch triển khai.*

The project goal is to implement several clustering models on an extensive image dataset and compare these methods. We provide a Web application to visual clustering results.

1. **Research scope and content / Nội dung nghiên cứu**

*Clearly define the key study topics and the project' scope. / Nêu rõ những nội dung nghiên cứu chính, phạm vi bao quát của đề tài.*

Key study topics:

* Clustering images using Bag of Visual Words and local feature descriptors
* Clustering using deep feature learning model.
* Evaluate the performance, drawbacks, and future developments.

Project' scope:

* Our research focuses on clustering methods, feature descriptors and feature reduction in the image processing field.
* The result obtained from this research can be used as the foundation for someone who wants to cluster a big image dataset

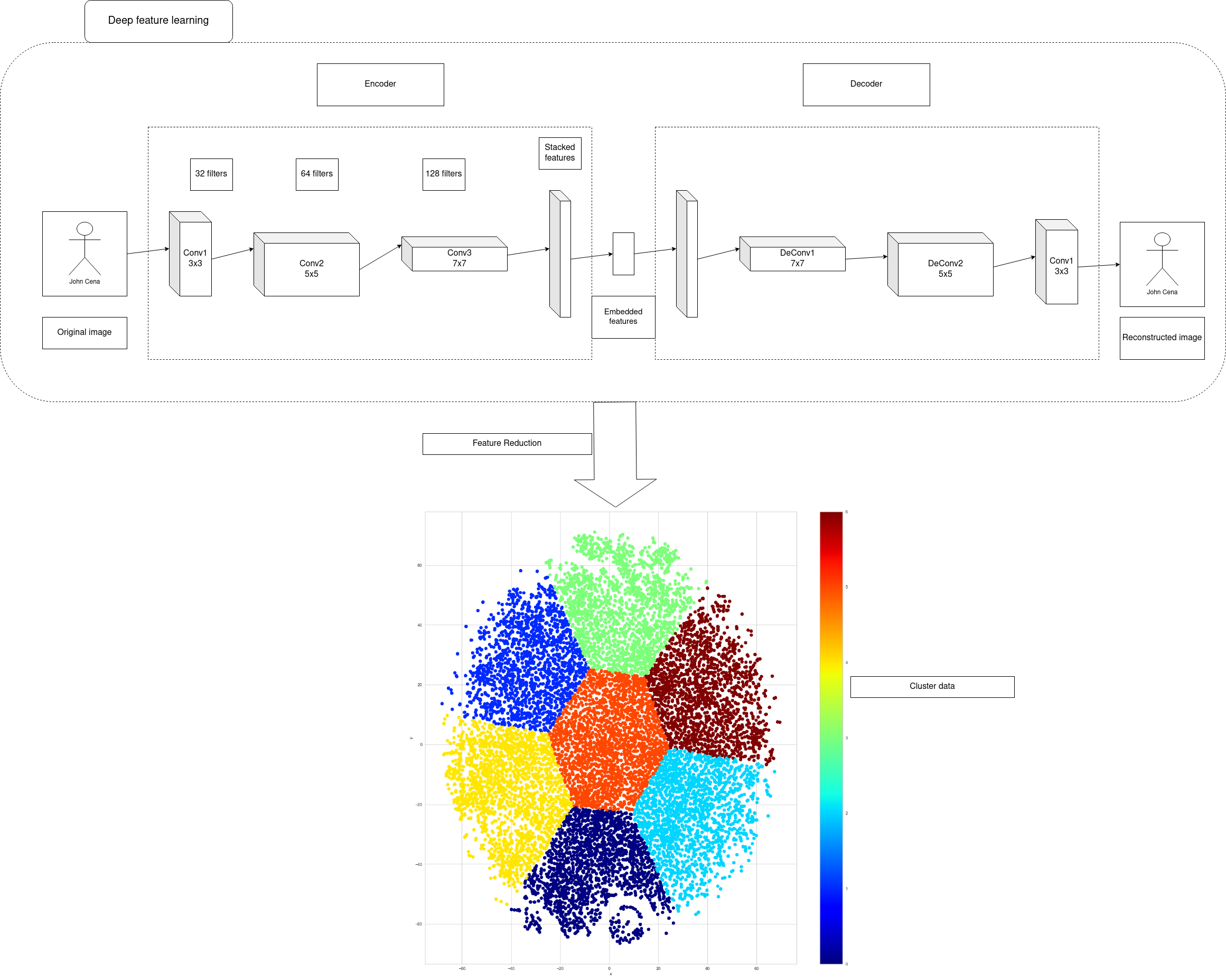
1. **Approach / Cách tiếp cận, phương pháp nghiên cứu**

*Describe research methodology and techniques appropriate for each issue of the project. / Mô tả chi tiết cách tiếp cận, phương pháp nghiên cứu, kỹ thuật sử dụng phù hợp với từng nội dung của đề tài nhằm đạt được mục tiêu của đề tài.*

Research is carried out based on experimental methods:

Research the current situation of clustering on image datasets. --> Make a compilation of some clustering models on image datasets --> Implement these models --> Software engineering

Figure 1 describes in detail the deep clustering model:



*Figure 1: Approach overview*

1. **Expected results / Dự kiến kết quả đề tài**

*For example: new theories, methods, applications; scientific significance and potential application, and plans to apply this research in the reality. / Nêu dự kiến kết quả nghiên cứu (phát hiện mới, lý thuyết mới, phương pháp mới, ứng dụng mới, …); ý nghĩa khoa học, khả năng sử dụng kết quả nghiên cứu, và kế hoạch áp dụng nghiên cứu trên thực tế.*

The research project will propose a comprehensive survey of different ways to cluster a big image dataset on speed and accuracy. We implement a model and build a software system to visualize the result.

**7. References**

[1]Rai, Pradeep & Shubha, Singh. (2010). A Survey of Clustering Techniques. International Journal of Computer Applications. 7. 10.5120/1326-1808.

[2] Mittal, H., Pandey, A.C., Saraswat, M. *et al.* A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimed Tools Appl* (2021). <https://doi.org/10.1007/s11042-021-10594-9>

[3] Li, F., Fergus, R. and Torralba, A., 2009. *Recognizing and Learning Object Categories*. [online] People.csail.mit.edu. Available at: <http://people.csail.mit.edu/torralba/shortCourseRLOC/> [Accessed 10 June 2021].

[4] C. Leng, H. Zhang, B. Li, G. Cai, Z. Pei and L. He, "Local Feature Descriptor for Image Matching: A Survey," in *IEEE Access*, vol. 7, pp. 6424-6434, 2019, doi: 10.1109/ACCESS.2018.2888856.

[5] Kavitha, B.R. & G, Ramya & Govindaraj, Priya. (2019). Performance comparison of various feature descriptors in object category detection application using SVM classifier. International Journal of Innovative Technology and Exploring Engineering. 8. 461-464.

[6] Rublee, Ethan & Rabaud, Vincent & Konolige, Kurt & Bradski, Gary. (2011). ORB: an efficient alternative to SIFT or SURF. Proceedings of the IEEE International Conference on Computer Vision. 2564-2571. 10.1109/ICCV.2011.6126544.

[7] J. L. Schönberger, H. Hardmeier, T. Sattler and M. Pollefeys, "Comparative Evaluation of Hand-Crafted and Learned Local Features," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6959-6968, doi: 10.1109/CVPR.2017.736.

[8] Luo, Juan & Oubong, Gwun. (2009). A comparison of sift, pca-sift and surf. International Journal of Image Processing. 3.

[9] Nasreen, Shamila. (2014). A Survey Of Feature Selection And Feature Extraction Techniques In Machine Learning, SAI, 2014.

[10] G. C. Nutakki, B. Abdollahi, W. Sun, O. Nasraoui (2019). An Introduction to Deep Clustering, DOI: 10.1007/978-3-319-97864-2\_4

[11] Y. Ren, N. Wang, M. Li, Z. Xu (2020). Deep density-based image clustering, DOI: 10.1016/j.knosys.2020.105841

[12] Z. Li, W. Yang, S. Peng, F. Liu (2020). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects, arXiv:2004.02806

[13] designbyhumans. <https://www.designbyhumans.com/>. [Accessed: 8-June-2021]