VISUAL HAND GESTURE RECOGNITION

Dr. N. Vinaya Kumari[1], G. Bala Xavier[2], B. Tarun3, M. Saketh[4],

[1]Associate Professor, [2, 3, 4,] Students
[1,2, 3, 4,5]Department of Artificial Intelligence and Machine LearningMallaReddy Institute of Technology and Science,

Hyderabad, India.

Email Id: v.vinayakumari@gmail.com, bjxavier662@gmail.com,
bhosletarun7321@gmail.com,sakethmacha123@gmail.com

I. ABSTRACT

Hand gesture recognition is a natural means of human-computer interaction (HCI) and has many potential applications, such as sign language recognition, virtual reality, and gaming. Convolutional neural networks (CNN) are a type of deep learning model that have proven to be very effective for image recognition tasks, including hand gesture recognition. Convolutional neural networks (CNN) have proven to be very effective in image recognition tasks, making them an ideal choice for hand gesture recognition. This study explores the development of a robust hand gesture recognition system using CNN. The proposed system starts by collecting a rich and diverse hand gesture dataset, including common signs, gestures, and movements. The dataset is pre-processed to improve image quality, remove noise, and extract relevant features. Data augmentation techniques are applied to increase the size of the data set and improve the generality of the model. The deep CNN architecture is designed and trained on a pre-processed dataset. The network consists of several convolutional layers, followed by pooling layers and fully connected layers. Transfer learning is also explored by fine-tuning pre-trained CNN models such as VGG16 or ResNet50, which has the potential to improve recognition accuracy. To evaluate the performance of the system, a comprehensive set of metrics including accuracy, precision, recall, and F1 score are used. Real-time gesture recognition is achieved by deploying the trained model to edge devices, ensuring low latency and efficient recognition. Experimental results demonstrate the effectiveness of the proposed CNN-based hand gesture recognition system. Achieve high precision and real-time performance, even in difficult lighting conditions and with a variety of hand sizes and orientations. Potential applications of the system include sign language interpretation, virtual reality control, and human-robot interaction, helping to improve accessibility and usability in a variety of fields. In this summary, we review the state of the art in hand gesture recognition using CNNs. We discuss different approaches to hand gesture recognition, the challenges involved, and the latest advances in the field. We also highlight some potential applicationsof hand gesturerecognition.

Keywords:

HCI- Human Computer Interaction CNN- Convolution

Neural Network VGG- Very Deep Convolution Network

ResNet50- 50 layer deep of Convolution Neural Network

II. INTRODUCTION

Handgesturerecognitionhasbeenapromisingtopicandappliedtomanypracticalapplications [1]. For example, hand gesture is observed and recognized by surveillance cameras to prevent criminal behaviours [2]. Also, hand gesture recognition has been investigated by a variety of studies[3].suchassighlanguagerecognition[4],liedetection[5],androbotcontrol[6].Fora image-based human hand gesture recognition system, since the number of variables of an image space is widely large, it is crucial to extract the essential features of the image. To implementagoodhandgesturerecognitionsystem,alargetrainingdatabaseisusuallyrequired andvariousgesturesshouldbemodelled.Withoutmucheffortonmodellingdifferentgestures, we develop a human gesture recognition system based on a Convolution Neural Network (CNN) in which the skin color model is improved and the hand pose is calibrated to increase recognition accuracies. CNNs can learn spatial and temporal features from hand gesture images, allowing them to accurately classify hand gestures even under difficult conditions, such as hand occlusion, posture variation, and lighttransformation.

Hand gesture recognition is a natural way for humans to communicate and interact with the world around them. It has many potential applications, such as sign language recognition, virtual reality, gaming, and assistive technologies. Convolutional neural networks (CNN) are atypeofdeeplearningmodelthathaveproventobeveryeffectiveforimagerecognitiontasks, including hand gesture recognition. CNNs can learn spatial and temporal features from hand gesture images, allowing them to accurately classify hand gestures even under difficult conditions, such as hand occlusion, posture variation, and lighttransformation.

III. LITERATURE SURVEY AND COMPARATIVEANALYSIS

Literature survey and comparative analysis are essential elements of research in the field of hand gesture recognition using convolutional neural networks (CNN). These activitiesinclude reviewingexistingliterature,research,andresearchmethodsrelatedtothetopicandcomparing them to identify trends, strengths, weaknesses, and gaps in knowledge. Gesture is a body language that humans use it to express emotion and thoughts. The varied gestures of the five fingersandpalmmayhavetheirphysicalmeanings.Handgesturerecognitionisacomplicated system that is composed of true modelling, gesture analysis and recognition, and machine learning.Inpreviousworkonmodellinggestures,HiddenMarkovModel(HMM)wasusedto areal-timesemanticlevelAmericanSignLanguagerecognitionsystem[7].Agesturealsocan be modelled as a HMM state sequence. 2014 IEEE International Conference on Automation Science and Engineering (CASE) Taipei, Taiwan, August 18-22, 2014 978-1-4799-5283- 0/14/$31.00 ©2014 IEEE 1038 In [8], they adopted a Finite State Machine (FSM) model to recognize human gestures. In [9], Time Delay Neural Network (TDNN) was used to match motion trajectories and train gesture models. Feature extraction plays an important role in a human gesture recognition system because the information about shape, pose, and texture of a gestureishelpful.Forexample,fingertips[10]andhandcontour[11]wereusedasthetraining features to build the gesture model. But the various light conditions have severe influences to gesture recognition because non-geometric features such as color, silhouette and texture are unstable. Using gesture semantic analysis is suitable for recognizing a sequence of gestures in doingacomplextask,butisinsufficienttocorrectlyrecognizegesturesinasimpletenuous motion. Jo, Kuno and Shirai [12] used FSM to deal a task-level recognition problem where a task was represented a state transition diagram and every state represented a possible gesture. Some researchers used a rule-based method for gesture recognition. Culter and Turk [13] designed a set of rules to recognize waving, jumping, and marching gestures. In recent years, deep learning is widely applied to many applications. Especially, CNN is a proper method for image-basedlearning.Forexample,[14]usedaCNNtoimplementrecognizeopenandclosed hands.

IV. METHODOLOGY

ThissectionprovidesthedescriptionofthedatasetandCNNconfigurationthatwereused.The flowchart of methodology is shown on Figure 1. The approach is the combination of data collection, pre-processing, configuring the CNN and building themodel.

A. Input Data and TrainingData

Imagesneededtotrainandvalidatethemodelwerecollectedusingawebcam.Thegestures were performed by 10 persons in front of the webcam. It is assumed that the input images exactly include one hand, gestures were made with right hand, the palm facing the camera andthehandwereroughlyvertical.Therecognitionprocesswillbelesscomplexandmore efficient if the background is less complex and the contrast is high on the hand. So, it is assumed that the background of the images was less complex anduniform.

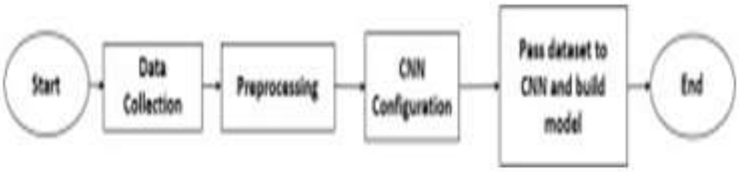

Fig. 1. System Framework

A minimal pre-processing was applied over the dataset

to reduce the computational complication and achieve better efficiency. Firstly, the background of the images was removed using the method of background subtraction proposed by Z. ZivKovic [15] [16]. The background subtraction is mainly based on K-gaussian distribution which selects appropriate gaussian distribution for each pixel and provides a better adaptability on varying scenes due to illumination changes. After subtracting background, only the image of hand remains. Then the images were converted to grayscaleimage.Sincegrayscaleimagescontainonlyonecolorchannelitwillbeeasier for CNN to learn [17]. Then a morphological erosion was applied [18]. After that, medianfilterwereappliedtoreducethenoise.Insignalprocessing,itisoftendesirable to reduce noises [19]. Figure 2 visualizes the pre-processing steps. The images were thenresizedtosize50x50forfeedingtoCNN.Inadditiontoourself-developeddataset, anotherdatasetnamed”HandGestureRecognitionDatabase”[20]wasalsousedinthis

experiment. By selecting largest object, hand in this case, other objects from these images were removed.

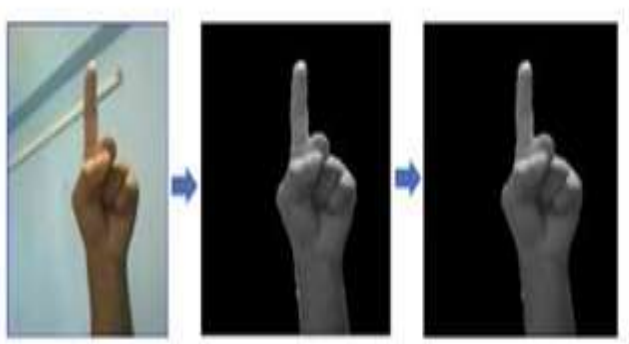

Fig. 2. Steps of Preprocessing

B. Dataset

We selected 10 static gestures (Index, Peace, Three, Palm Opened, Palm Closed, OK, Thumbs, Fist, Swing, Smile) to recognize. Each class has 800 images for training and 160 images for testing purpose. So total number of images is 8000 .

(a)

(b)

for training and 1600 for testing. Sample of finalized dataset is provided on Figure 4. ”HandGestureRecognitionDatabase”[21]alsocontains10classes(Palm,I,Fist,Fist Moved, Thumb, Index, OK, Palm Moved, C, Down), each class having 2000 images. A snapshot from the database is provided in Figure5.

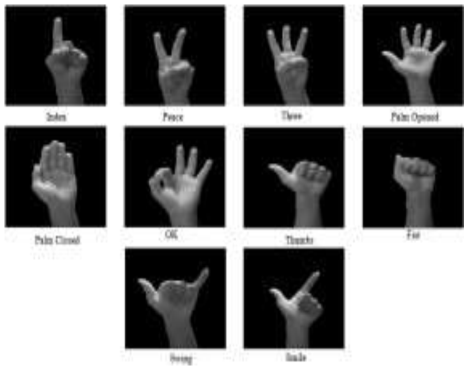

Fig. 4. Sample Images from self-developed Dataset

C. CNNConfiguration

The CNN that that has been considered in this research to recognize hand gesture is composed of two convolution layers, two max pooling layers, two fully connected layers and layer. There are three dropout performance in the network to prevent over-fitting[22]. The first convolution layer has 64 different filters with the kernel size 3x3. The activation functionusedinthislayerisRectifiedLinearUnit(ReLU).ReLUwasappliedtointroduce non-linearity [23] and it has been proved that ReLU performs better than other activation functions such as tanh or sigmoid. As it is input layer, we have to specify the input size. The stride is set to default. The input shape is 50x50x1 which means that gray-scaleimage ofsize50x50shouldbeprovidedtothisnetwork.Thislayerproducesthefeaturemapsand passes them to the next layer. Then the CNN has a max pooling layer with pool size 2x2 which takes the maximum value from a window of size 2x2. The spatial size of the representationisreducedprogressivelyasthepoolinglayertakesonlythemaximumvalue and discards the rest. This layer helps the network to understand the images betterbecause it only selects more important features. The next layer is another convolution layer and it has 64 different filters with the kernel size 3x3 and defaultstride.

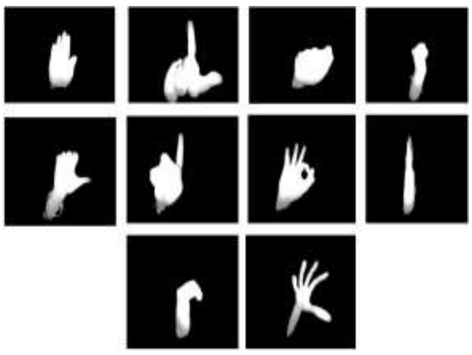

Fig. 5. Sample Images from Hand Gesture Recognition Database

Again, ReLU was used as the activation function in this layer. This layer is followed by another max pooling layer which has a pooling size 2x2. In this layer, first dropout was added which randomly discards 25% of the total neurons to prevent the model from over-fitting.Outputfromthislayerispassedtotheflattenlayer.Outputfromthepreviouslayers are received by the flattening layer and they are flattened to a vector fromtwo-dimensional matrix. This layer allows the fully connected layers to process the data achieved till now. The next layer is first fully connected layer which has 256 nodes and ReLU was used as the activation function. The layer is followed by a dropout layer which excludes 25% of the neurons to prevent overfitting. The second fully connected layer again has 256 nodes to receive the vector produced by first fully connected layer and uses ReLU as activation layer. The layer is followed by a dropout layer to exclude 25% of the neurons to prevent overfitting. The output layer has 10 nodes corresponding to each classes of the hand gestures. This layer uses SoftMax function [24] as activation function which outputs a probabilistic value for each of the classes. The model is then compiled with Stochastic Gradient Descent (SGD) [24] function with a learning rate 0.001. To evaluate loss, categoricalcross-entropyfunction[25]wasusedsincethemodeliscompiledformorethan two classes. Finally, the metrics of loss and accuracy were specified to keep track on the evaluation process. This configuration was chosen after trying various combination of nodes andlayers.

D. SystemImplementation

Toimplementthesystem,pythonwasusedastheprogramminglanguageandapythonIDE Spyder was used to write and run code. The library Keras was used for building the CNN classifier.ThelibraryPILwasusedforimagepreprocessing.Sklearnwasusedtocalculate theconfusionmatrix.Matplotlibwasusedtovisualizemodelaccuracyandlossvaluesand confusionmatrix.NumPywasusedforarrayoperations.Thetrainingprocessondatasetis composed of two phases. 1) Training with Base Dataset: In this phase, the model was trained using the base dataset achieved after pre-processing. 2) Training with Expanded Dataset: In this phase, the dataset was augmented. Data augmentation is a technique to increase the number of data by applying zoom, shear, rotation, flip etc [26]. This process not only increases the data but also brings variation in dataset which is essential for CNN to learn sophisticated differences of images. A random image was selected to provide the demonstration

V. RESULT AND DISCUSSION

Theresultsofhandgesturerecognitionusingconvolutionalneuralnetworks(CNN)havebeen very promising in recent years. CNN-based hand gesture recognition systems have achieved state-of-the-art results on many complex hand gesturedatasets.

Forexample,theHandGestureTransform(HGT)architectureachieves99.8%accuracyonthe AmericanSignLanguage(ASL)fingerspellingdataset.Dual-streamCNNalsoachievesstate- of-the-art results on many hand gesture datasets, with accuracy typically above95%.

CNN's success in hand gesture recognition is due to its ability to learn spatial and temporal features from hand gesture images. Spatial features relate to the arrangement of pixels in an image, while temporal features relate to changes in the image over time. CNNs are capable of learning spatial and temporal features using a series of convolutional and pooling layers.

CNN-based hand gesture recognition systems have many advantages compared to traditional hand gesture recognition systems. CNN-based systems are more robust to manual masking, pose changes, and illumination changes. They can also learn complex hand gesture patterns. Start by presenting quantitative results of your CNN model's performance. Include precision, accuracy, recall, F1 score, and any other relevant evaluation metrics. Use tables, graphs, or charts to make results visually accessible. Show a confusion matrix to illustrate how well your model classifies different hand gestures. This can help determine which gestures are correctly recognized and where the model may be having difficulty. Compare with baseline:If youhave conducted experiments with multiple models (including base models or variations of theCNN architecture), compare their performance. Highlight any significant improvements or compromises.Explaintheaccuracyofyourmodel.Discusstheoverallrecognitionrateofhand gestures. Correct any differences in accuracy between different gesture classes. Discuss the challenges your model faces. This can include issues with lighting conditions, background clutter, variations in hand size or orientation, and gestures that looksimilar.

If your model is deployed for real-time recognition, discuss its latency and efficiency. Refers to any optimizations or compromises made to achieve real-time performance.If you apply transfer learning with pre-trained models, discuss how this affects your results. Show whether this improves recognition accuracy and reduces training time. Consider any potential bias in your data set. Discuss how the diversity (or lack of diversity) of the data set can impact the generalization of the model to real-world situations[10].Discuss the practical implications of your results. How can your model benefit users in real-world applications, such as sign language translation, gaming, or human-computer interaction? Future orientation: Identify areas that require additional research or improvement based on the limitations and challenges you face. Suggest potential solutions or approaches to address these problems. If you perform a comparative analysis with existing research, discuss how your results are consistent or differentfrompreviousstudies.Highlightanyuniquecontributionsyourresearchbringstothe field.

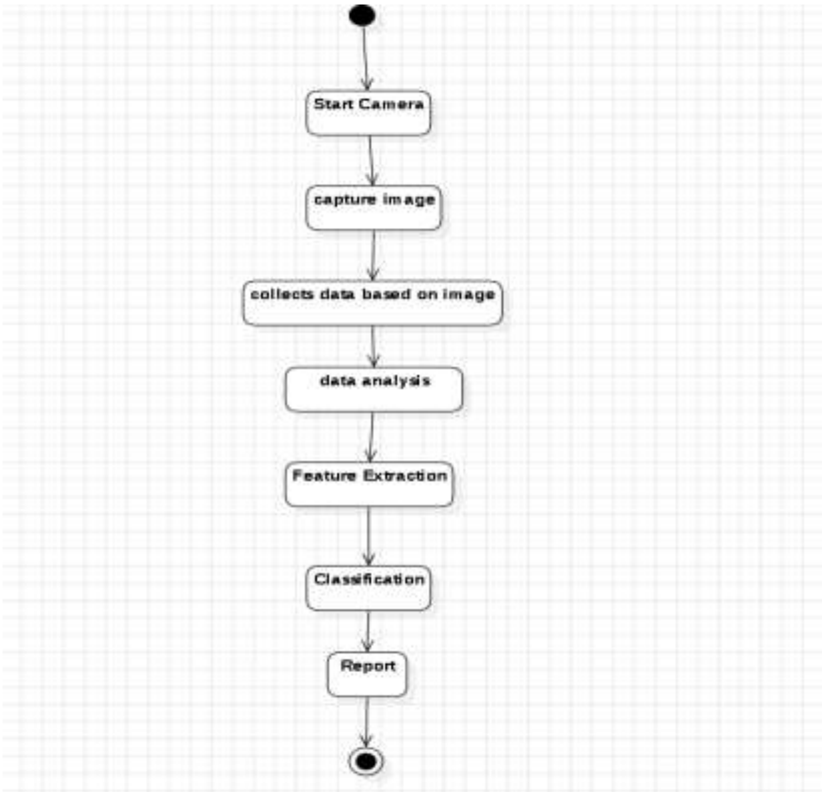

Figure 01: Activity Diagram of Hand gesture Recognition

VI. CONCLUSION AND FUTURESCOPE

Conclusion:

Hand gesturerecognition using convolutional neuralnetworks (CNN) is a rapidly growing field with many potential applications. CNN-based hand gesture recognition systems have achieved state-of-the-art results on many complex hand gesture datasets.

CNN-based hand gesture recognition systems have many advantages compared to traditional hand gesture recognition systems. CNN-based systems are more robust to manual masking, posechanges, and illumination changes. They can also learn complex hand gesture patterns. CNN-based hand gesture recognition systems have the potential to revolutionize the way we interact with computers and the environment. They can be used to develop new and innovative applications in many fields, such as sign language recognition, virtual reality, gaming and assistive technology.

Futurescope:

- There is still significant room for improvement in hand gesture recognition using CNN. Some promising future directions in this field include:
- Develop a more robust and accurate hand gesture recognition system capable of operating in real time and under challenging conditions.
- Develop new and innovative hand gesture recognition applications in more fields.
- Explore the use of other types of deep learning models for hand gesture recognition, such as recurrent neural networks (RNN) and short-term memory networks (LSTM).
- Developed new data augmentation techniques to improve the performance of CNN- based hand gesture recognition system on limited training data.

I believe that hand gesture recognition using CNN is a very promising field that can have a great impact on our lives. I look forward to seeing how this technology develops in the coming years.

VII. ACKNOWLEDGMENT

The team members of the research project want to sincerely thank our guide Associate Professor Dr. Vinaya Kumari and the Department of Computing Science and Engineering, Malla Reddy Institute of Technology and Science, India for their encouragement and support for completion of this work.

VIII. REFERENCES

[1]. A. Kojima, M. Izumi, T. Tamura, and K. Fukunaga, "Generating natural language description of human behaviour from video images" in Int. Conf. Pattern Recog. , vol. 4. IEEE, 2000, pp. 728–731.

[2]. C. J. Cohen, F. Morelli, and K. A. Scott, "A surveillance system for the recognition of intent within individuals and crowds," in IEEE. Conf. Technol. For Homeland Secur. IEEE, 2008, pp. 559–565.

[3]. S. Mitra and T. Acharya,"Gesture recognition: A survey," IEEE Trans. Syst. Man, Cybern. C, vol. 37, no. 3, pp. 311–324, 2007.

[4]. C. Vogler and D. Metaxas, "ASL recognition based on a coupling between hmms and 3d motion analysis" in Int. Conf. Computer Vision. IEEE, 1998, pp. 363–369.

[5]. C. F. Bond Jr, A. Omar, A. Mahmoud, and R. N. Bonser, "Lie detection across cultures"J. nonverbal behave. , vol. 14, no. 3, pp. 189–204, 1990.

[6]. H. I. Lin, C. H. Cheng, and W. K. Chen, "Learning a pick-and-place robot task from human demonstration," in Proc. Int. Conf. Automat. Control. IEEE, 2013, pp. 312–317.

[7]. T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer-based videos" IEEE Trans. Pattern Anal. Mach. Intell, vol. 20, no. 12, pp. 1371– 1375, 1998.

[8]. J. Davis and M. Shah, "Visual gesture recognitions" in IEE Proc. Vision, Image and Signal Process. , vol. 141, no. 2. IET, 1994, pp. 101–106.

[9]. M. -H. Yang and N. Ahuja, "Recognizing hand gestures using motion trajectories," in Face Detection and Gesture Recognition for HumanComputer Interaction. Springer, 2001, pp. 53– 81.

[10]. . K. Oka, Y. Sato, and H. Koike, "Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems" in IEEE Int. Proc. Automat. Face and Gesture Recog. IEEE, 2002, pp. 429–434.

[11]. A. A. Argyros and M. I. A Lourakis, "Vision-based interpretation of hand gestures for remote control of a computer mouse," in Computer Vision in Human-Computer Interaction. Springer, 2006, pp. 40–51.

[12]. K. -H. Jo, Y. Kuno, and Y. Shirai, "Manipulative hand gesture recognition using task knowledge for human computer interaction," in IEEE Int. Conf. Automat. Face and Gesture Recog. IEEE, 1998, pp. 468– 473.

[13]. R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," in IEEE Int. Conf. and Workshops on Automat. Face and Gesture Recog. IEEE Computer Society, 1998, pp. 416–416.

[14]. S J. Nowlan and J. C. Platt, "A convolutional neural network hand tracker," Advances in Neural Inf. Process. Systems, pp. 901–908, 1995.

[15]. Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in null. IEEE, 2004, pp. 28–31.

[16]. Z. Zivkovic and F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," Pattern recognition letters, vol. 27, no. 7, pp. 773–780, 2006.

[17]. M.Grundland and N. A Dodgson, "Decolorize: Fast, contrast enhancing, color to grayscale conversion" Pattern Recognition, vol. 40, no. 11, pp. 2891–2896, 2007.

[18]. R.M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology" IEEE transactions on pattern analysis and machine intelligence, no. 4, pp. 532– 550, 1987.

[19]. Y.Zhu and C.Huang, "An improved median filtering algorithm for image noise reduction," Physics Procedia, vol. 25, pp. 609–616, 2012.

[20]. T. Mantecon, C. R. del Blanco, F. Jaureguizar, and N. Garc "Hand gesture recognition using infrared imagery provided by leap motion controller " in International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, 2016, pp. 47–57.

[21]. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting" The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[22]. X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 315–323.

[23]. R. A. Dunne and N. A. Campbell, "On the pairing of the softmax activation and cross- entropy penalty functions and the derivation of the softmax activation function" in Proc. 8th Aust. Conf. on the Neural Networks, Melbourne, vol. 181. Citeseer, 1997,pp. 185.

[24]. L. Bottou "Large-scale machine learning with stochastic gradient descent" in Proceedings of COMPSTAT'2010. Springer, 2010, pp. 177– 186.

[25]. L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv: 1712. 04621, 2017.