Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

# Enhancing Classification in Utd-mhad Dataset: Utilizing Recurrent Neural Networks in Ensemble-based Approach for Human Action Recognition

Saketh Kilaru
Electronics Department
BITS Pilani
Pilani, India

Anushka Shah
Electronics Department
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Swoichha Adhikari
Software Department
Gandaki College of Engineering and Science
Pokhara, Nepal

***Abstract*—**This study demonstrates the ensemble approach to perform Human Action Recognition on the UT Dallas' Multimodal Human Action Data, where number of actions by humans is 27. Our ensemble approach gained an accuracy of 0.821 on the validation data, a remarkable uplevelling as compared to the accuracy of baseline paper which is 0.672. The paper also shows the train-val performances of other models we experimented using only the Inertial and Skeleton dataset. The link to Github repository which holds to code can be found here, and the link to get the dataset can be found here.

## I. INTRODUCTION

Human Action Recognition(HAR), a research domain that piqued interests from diverse range of computer science disciplines from the period of the 1980s, as a result of its application in numerous branches of investigstion like sociology, human-computer interaction, and medicine. Data which is in the format of videos, inertial sensors, like accelerometers and gyroscopes, depth maps and point clouds are often used to classify the different human actions [1][2]. Traditional approaches are generally broken down as the following three steps1) Feature Extraction using Signal Processing or Computer Vision techniques to capture relevant spatio-temporal features [3][4][5] 2) Designing a pipeline to combine the extracted features 3) Training a classifier, usually a Support Vector Machine(SVM) or Random Forest, using following charcteristics The mentioned approach often requires deep domain knowledge in extracting the useful spatio-temporal features. Another approach is to use modern neural network architectures to do the feature extraction and model building automatically. Soon after 2014, there were two breakthrough papers, Single Stream and Two Stream, which ignited the research using modern approach. The main difference between them was how the model architecture combines the spatiotemporal information. The first paper [6] uses a Single Stream Network(SSM) which explores distinct pathways to club temporal info and consecutive frames by utilizing 2D pre-trained convolutions. This had led to popular methods like Longterm Recurrent Convolutional Networks[7], 3-D Convolutional Networks [8], and Conv3D with Attention [9]. The second paper utilizes a Two Stream Network [10], as depicted in Figure 1, one stream captures the pre-trained spatial context, while the second one captures motion information. Other variants are soon developed as like the Two-Stream Network Fusion [11], Temporal Segment Networks (TSN) [12], Action Video-Level Aggregation (ActionVLAD) [13], Hidden TwoStream Convolutional Networks [14], Two-Stream Inflated 3D ConvNet (TwoStreamI3D) [15], and Temporal 3D ConvNets (T3D) [16] Fig. 1.

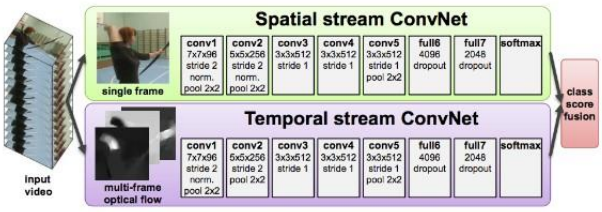


Fig. 1. 2 stream architecture of Simmoyan and Zisserman [10]

This paper presents an ensemble proposition, using both convolutional as well as recurrent, single stream neural networks to recognize 27 different human actions using the UTDMHAD dataset [17].

## II. METHODS AND MODELING

### A. Dataset

Derived through the integration of diverse technologies, a wearable inertial sensor like Microsoft Kinect sensor equipped with an accelerometer and gyroscope, and a video camera, the UTD-MHAD dataset encompasses 27 unique actions executed by eight individuals (four females and four males). All of them performed each action four times for data collection purposes. Following the elimination of three faulty sequences, and there are 861 sequences in totality. There are four different types of datasets, the Depth dataset, Inertial dataset, Skeleton dataset, and RGB dataset. They are plotted and shown respectively in Fig. 2 - 3.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

Fig. 2. Depth map dataset of a one person doing a tennis swing

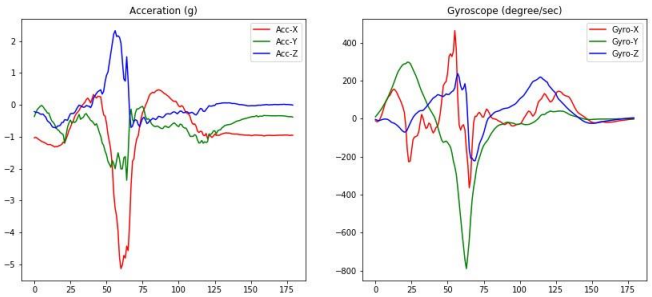


Fig. 3. Inertial dataset of a one person doing a tennis swing

In this paper, we would only be using the inertial and skeleton dataset for our models. We split the train-validation data in the same fashion as the original paper [17],For training purposes, the dataset will utilize subjects 1, 3, 5, and 7, while validation will involve subjects 2, 4, 6, and 8, so that we could perform a baseline comparison.

B. Data Pre-Processing

Inertial Data: The inertial data was re-sampled to the mean period of 180 units. This was found to allow the model to achieve convergence much quicker. Amplitude normalization was also tried but subsequently removed as it does not show improvement in the model's training. 2. Skeleton Data: To aid with data fusion later, we resampled the skeleton data having 180 frames.

C. Models

During our experiments, we employed the AdamOptimizer with the following values: = 1e 4, 1 = 0.9, and 2 = 0.999. To initialize the model's weights, we utilized the Glorot initialization method. To improve generalization, the batch size was set to 3, as suggested in [18]. In this paper, we shall experiment with the following models:

1. Simple LSTM
2. BiDirectional LSTM
3. Conv LSTM
4. UNet LSTM
5. Ensemble of Conv LSTM and UNet LSTM

The full architecture of the models can be found in the Appendix. Simple LSTM: We start off by having a Simple LSTM model, which comprises of an LSTM unit of 512 hidden units, followed by a Dense layer, to classify all 27 actions. Our remarkable strategy enables rapid handling of regional, decentralized, continuous numerical, and distorted structure representations, surpassing conventional real-time iterative learning, time-driven error optimization, sequential cascade interdependence, Elman networks, and neural sequence segmentation techniques in terms of both speed and efficiency. Each recurrent unit in the LSTM was set to have a dropout rate of 0.2.

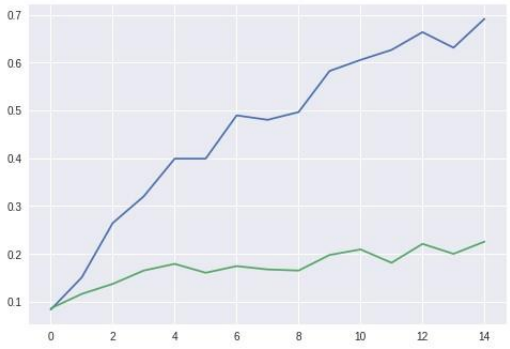

Fig. 4. Train (blue)-val (green) accuracy plot of the Simple LSTM model

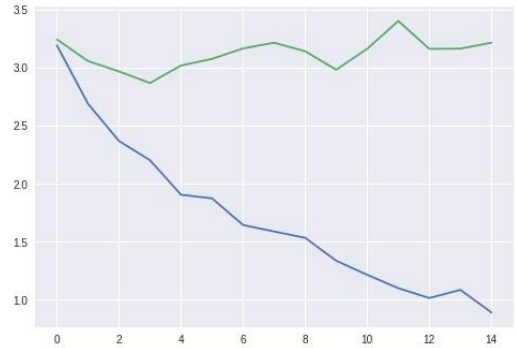

Fig. 5.Train (blue)-val (green) loss plot of the Simple LSTM model

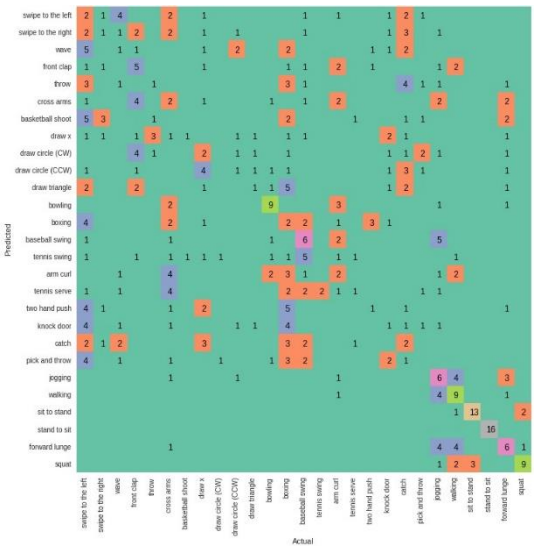


Fig. 6. The assessment of the Simple LSTM model's performance is carried out using a confusion matrix.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

BiDirectional: The next model flips the copy of LSTM unit and concatenates it with original LSTM unit. With our groundbreaking technique, we realize a variant of the process of generating deep learning, allowing the layer of output to gather insights pertaining to both past as well as future possibility of states at the same time [19]. Afterward, a fully connected layer with a softmax activation is utilized for classifying the 27 activities. This novel process enhances overall performance and capabilities of the model.

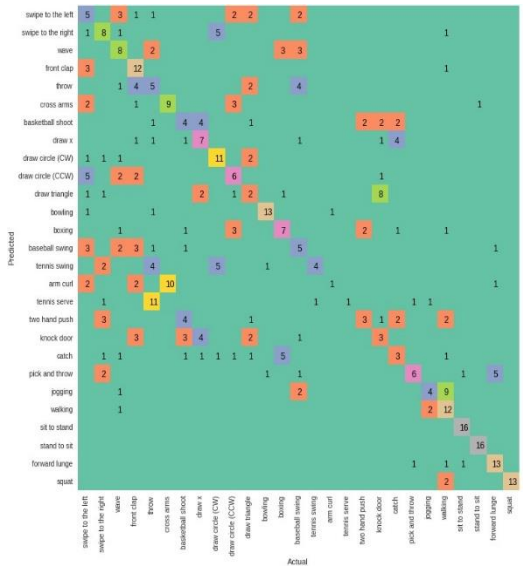


Fig. 7. The LSTM model with bidirectional connections confusion matrix is used for performance evaluation.

Conv LSTM: By employing a unique combination of 1D Convolutional and 1D Maxpooling layers, the third model efficiently extracts higher dimensional features. Following data processing, the information is input splitting into two LSTM units to note crucial corporeal details. Afterward, the LSTM unit's output undergoes flattening, while an added Dropout layer with a dropout rate of 0.5 boosts generalization capabilities. Finally, to achieve a robust and accuracy, we integrate a fully activated layer using softmax to enhance the model's classification capabilities allowing the classification of all 27 actions.

UNet LSTM: Incorporating a renowned architecture commonly utilized in image semantic segmentation roles, last model is founded upon the UNet framework.The UNet, a kind of convolutional neural network with encoder-decoder architecture, possesses symmetrical features in both its contraction and expansion paths. During the contraction phase, the input experiences multiple convolutions and maxpooling operations, resulting in amplified feature maps and diminished image resolution, favoring" what" over "where." During the expansion path, low-resolution, highdimensional features are up-sampled using convolutional kernels, resulting in reduced feature maps. Remarkably, UNet exhibits a distinctive integration of highdimensional characteristics from the contraction phase into the compact feature representations within the expansion layers, creating a robust and cohesive model.

Ensemble of Conv LSTM and UNet LSTM: The Conv LSTM and UNet LSTM were found to be performing well on the validation set, so we took an average of their softmax activation to create an ensemble. (1)

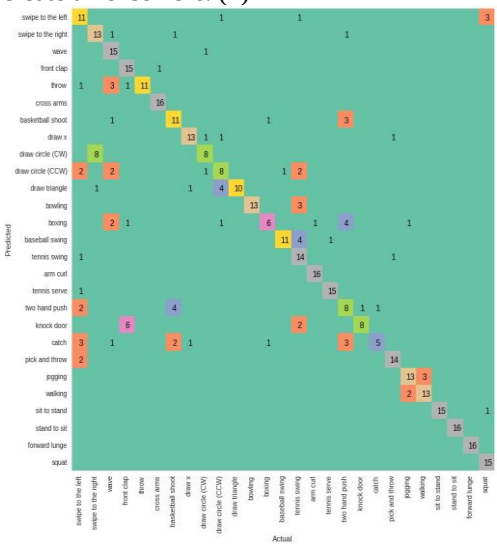


Fig. 8. The performance evaluation of the Ensemble of Conv LSTM and UNet LSTM model is represented using the confusion matrix.

## III. COMPUTATIONAL STIMULATION

The code is written in Python, using Keras with Tensorflow backend, NumPy, SciPy and Matplotlib libraries. The link to code's github repository is: https://github.com/notha99y/Multimodal human actions. The models are trained on Google Colaboratory notebooks.

A. Inertial Data Simple LSTM

The accuracy on the validation of our simple LSTM was 0.238. The train-val accuracy and loss plots of the Simple LSTM are shown in Fig. 11, 12 respectively. It shows that the model is quickly over-fitting after epoch 2. Bi-Directional LSTM: Similarly, the train-val accuracy, loss plots respectively, with the confusion matrix with a validation accuracy of 0.465. Conv LSTM: The Conv LSTM was the first major breakthrough

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

among our models, achieving a validation accuracy of 0.700. Lastly, our UNet LSTM got the highest accuracy of 0.712.

B. Inertial + Skeleton Data

Conv LSTM: We combined the Skeleton data along the features axis with the Inertial data and trained the Conv LSTM to achieve an accuracy of 0.784. The train-val accuracy, loss plots together with the confusion matrix is shown below.

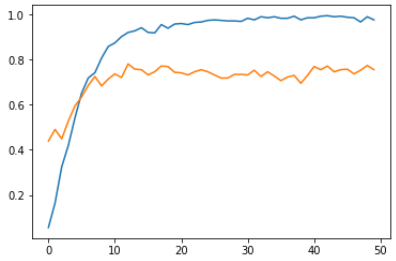

Fig. 9. Train (blue)-val (green) accuracy plot of the Conv LSTM model on both Inerital and Skeleton Data

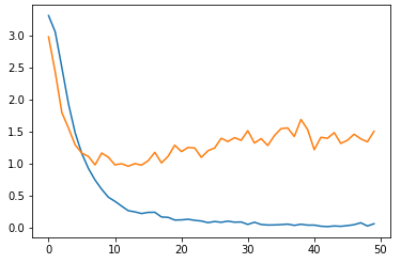

Fig. 10. Train (blue)-val (green) loss plot of the Conv LSTM model on both Inerital and Skeleton Data

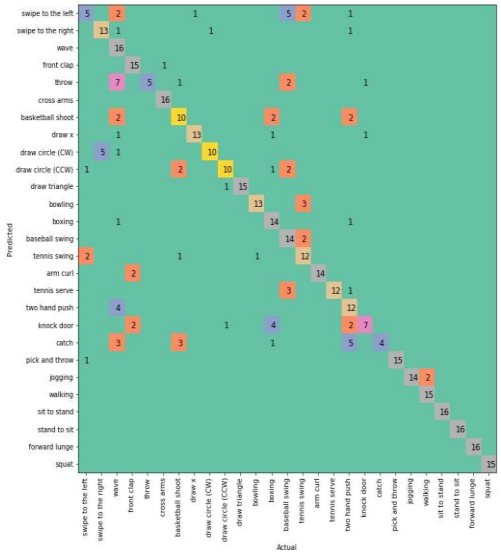


Fig. 11. Confusion matrix of the Conv LSTM model on both Inerital and Skeleton Data

2. UNet LSTM: Similiarly, the same was done with the UNet LSTM model and it achieve an accuracy of 0.742. 3. Ensemble of Conv LSTM and UNet LSTM: Lastly, we combined all our stocks together and put together an ensemble which achieved an accuracy of 0.821. The confusionn matrix.

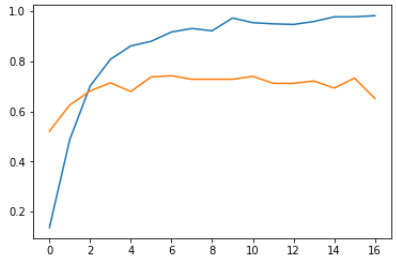

Fig. 12. Train (blue)-val (green) accuracy plot of the UNet LSTM model on both Inertial and Skeleton Data

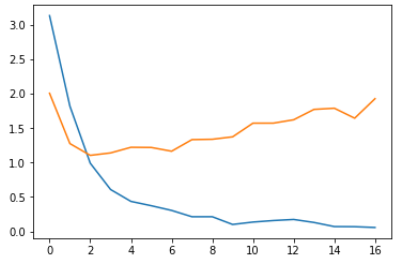

Fig. 13. Train (blue)-val (green) loss plot of the UNet LSTM model on both Inertial and Skeleton Data

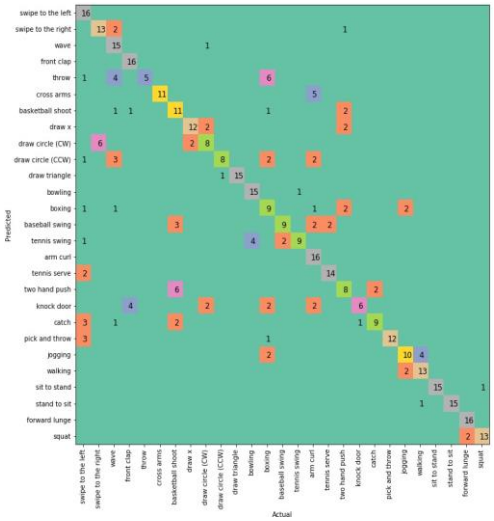


Fig. 14. The UNet LSTM model's confusion matrix is computed for both Inertial and Skeleton Data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

C. Summary

A summary of results can be found below in Table I.

TABLE I
AN OVERVIEW OF THE VALIDATION ACCURACY ACHIEVED BY VARIOUS MODELS.

| | S-LSTM | B-LSTM | C-LSTM | U-LSTM | Ensemble |
|---|---|---|---|---|---|
| Iner | 0.238 | 0.465 | 0.700 | 0.712 | 0.765 |
| Iner + skel | - | - | 0.784 | 0.742 | 0.821 |

## IV. CONCLUSION-DISCUSSION

This research showcases our ensemble technique in Human Action Recognition (HAR), resulting in a notable accuracy of 0.821 on the validation set, outperforming the baseline paper's results of 00.672, by utilizing Inertial and Skeleton Data. This could be due to the convolutional networks being able to capture more generic, higher dimensional features compared to the Collaborative Representation Classifier (CRC) method used in [4] MHAD. However, we find that there are still some things we can work on and in the following sub sections, we shall explore other methods as future work to improve our validation accuracy.

A. Issue of model overfitting

Throughout our endeavors, a persistent challenge was the early epoch over-fitting of our model. We effectively mitigated this issue by incorporating dropout layers, which encouraged exploration in alternative solution spaces by ensembling the UNet LSTM with the Conv LSTM [20]. Over-fitting typically arises when the model attempts to learn highfrequency features that offer little utility. To enhance the learning capacity, we introduced Gaussian Noise with a mean value of zero and data elements spanning across the entire spectrum. Additionally, we observed substantial variability in time sequences among different subjects, even for similar activities. To address this, we implemented augmentation of data through time-scaling, as well as translation, bolstering volume of data to be trained and enabling better generalization. It is pertinent to consider reducing the model's complexity to further mitigate overfitting risk while maintaining optimal performance.

B. Integration of Depth and RGB Data for Data Fusion

Combining Depth with RGB data in our fusion approach would lead to an increased number of input variables for training the models, resulting in enhanced validation accuracy.

C. Collective Intelligence

Right now, our ensemble simply takes the average of our models' softmax activations. We could further enhance the validation accuracy and reduce over-fitting, by exploring different Methods of Collaborative Learning[21]

## REFERENCES

[1] T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, P. . Klasnja, K. Koscher, A. LaMarca, J. A. Landay, L. LeGrand, J. Lester, A. Rahimi, A. Rea, and D. Wyatt, "The mobile sensing platform: An embedded activity recognition system," IEEE Pervasive Computing, vol. 7, no. 2, pp. 32–41, April 2008.

[2] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] S.-M. Lee, S. M. Yoon, and H. Cho, "Human activity recognition from accelerometer data using convolutional neural network," in 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Feb 2017, pp. 131–134.

[4] Laptev and Lindeberg, "Space-time interest points," in Proceedings Ninth IEEE International Conference on Computer Vision, Oct 2003, pp. 432–439 vol.1.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Oct 2005, pp. 65–72.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. FeiFei, "Large-scale video classification with convolutional neural networks," in CVPR.

[7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," CoRR, vol. abs/1411.4389, 2014. [Online]. Available: http://arxiv.org/abs/1411. 4389

[8] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," CoRR, vol. abs/1412.0767, 2014. [Online]. Available: http://arxiv.org/abs/1412.0767

[9] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing Videos by Exploiting Temporal Structure," ArXiv e-prints, Feb. 2015.

[10] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," CoRR, vol. abs/1406.2199, 2014. [Online]. Available: http://arxiv.org/abs/1406.2199

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional twostream network fusion for video action recognition," CoRR, vol. abs/1604.06573, 2016. [Online]. Available: http://arxiv.org/abs/1604. 06573

[12] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," CoRR, vol. abs/1608.00859, 2016. [Online]. Available: http://arxiv.org/abs/1608.00859

[13] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. C. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," CoRR, vol. abs/1704.02895, 2017. [Online]. Available: http://arxiv.org/abs/1704.02895

[14] Y. Zhu, Z. Lan, S. D. Newsam, and A. G. Hauptmann, "Hidden two-stream convolutional networks for action recognition," CoRR, vol. abs/1704.00389, 2017. [Online]. Available: http://arxiv.org/abs/1704. 00389

[15] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," CoRR, vol. abs/1705.07750, 2017. [Online]. Available: http://arxiv.org/abs/1705.07750

[16] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. V. Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," CoRR, vol. abs/1711.08200, 2017. [Online]. Available: http://arxiv.org/abs/1711. 08200

[17] C. Chen, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," 09 2015.

[18] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 1729–1739.

[19] M. Schuster, K. K. Paliwal, and A. General, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, 1997.

[20] R. Maclin and D. W. Opitz, "Popular ensemble methods: An empirical study," CoRR, vol. abs/1106.0257, 2011. [Online]. Available: http://arxiv.org/abs/1106.0257

[21] C.Zhang and Y.MA, Ensemble Machine Learning ,Methods and Applications. Landon: Springer, 2012, pp.254-270.

[22] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," 1997.

[23] R. E. Schapire, The strength of weak learnability. Kluwer Academic Publishers, 1990, p. 197-227.