IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

Emotion Recognition using Deep Learning

Ms. S. Maheshwari, Dr. R. Bhuvana, Ms. S. Sasikala
Assistant Professor, Department of Computer Science
Agurchand Manmull Jain College, Meenambakkam, Chennai, India
maheshwari.s@amjaincollege.edu.in, bhuvana.r@amjaincollege.edu.in, sasikala.s@amjaincollege.edu.in

Abstract: *Speech Emotion Recognition (SER) is critical in Human computer engagement (HCI) because it provides a deeper knowledge of the situation and leads to better engagement. Various machine learning and Deep Learning (DL) methods have been developed over the past decade to improve SER procedures. In this research, we evaluate the features of speech then offer Speech Former++, a comprehensive structure-based framework for paralinguistic speech processing. Following the component relationship in the speech signal, we propose a unit encoder to efficiently simulate intra- and inter-unit information (i.e., frames, phones, and words). We use merging blocks to generate features at different granularities in accordance with the hierarchy connection, which is consistent with the structural structure in the speech signal. Rather than extracting spatiotemporal information from hand-crafted features, we investigate how to represent the temporal patterns of speech emotions using dynamic temporal scales. To that end, we provide Temporal-aware bI- direction Multi-scale Network (TIM-Net), a unique temporal emotional modelling strategy for SER that learns multi-scale contextual affective representations from different time scales. Unweighted Accuracy (UA) of 65.20% and Weighted Accuracy (WA) of 78.29% are accomplished using signal features in low- and high-level descriptions, as well as various deep neural networks and machine learning approaches.*

Keywords: Human computer engagement, Deep Learning, Paralinguistic, Multi-scale Network, Weighted Accuracy

I. INTRODUCTION

Human emotional state is a significant aspect in their interactions, influencing most communication methods such as facial expressions, voice characteristics, and spoken language content [1–3]. One of the most common ways to convey emotions is through speech. It is critical to recognise, interpret, and respond to the emotions presented in speech in order to achieve natural human computer interaction (HCI) [3], [4]. Today, speech emotion recognition (SER) systems have a variety of applications, including natural human-machine interactions, such as web videos, computer videos, and training programmes, car driver safety, computer games, disease diagnostic tools, as a tool for an automatic translation system, and mobile communications [1]-[3].

Modelling speech signals in PSP is a difficult task since humans understand the pronunciation information and the dynamic changes in speech but models do not. Several machine learning techniques, such as hidden Markov models [2]- [4], decision trees [5], [6], and limited Boltzmann machines [7]-[9], have been proposed over the last three decades to capture paralinguistic information in speech. Deep learning algorithms have recently given higher performance for PSP problems due to their exceptional modelling capabilities. Convolutional neural networks (CNNs) [10–16], graph neural networks (GNNs) [17–18], recurrent neural networks (RNNs) [19– 20], and two popular RNN variants named long short term memory (LSTM) [12–14] and gated recurrent units (GRUs) [15] have all shown promising results in PSP.

Various temporal modelling approaches, in instance, Long Short-Term Memory (LSTM), Gate Recurrent Unit (GRU), and Temporal Convolution Network (TCN) are some examples. commonly used in SER, with the goal of capturing dynamic temporal fluctuations in voice signals. Wang et al. [7], for example, presented a dual-level LSTM to exploit temporal information from distinct time-frequency resolutions. Zhong et al. [9] learned integrated spatiotemporal characteristics using CNN with Bi-GRU and focus loss. Rajamani et al. [6] developed an attention- based ReLU within GRU to capture long-range feature interactions. Zhao et al. [8] learned spatiotemporal characteristics using completely

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

CNN and Bi-LSTM. These techniques, however, have the following drawbacks: 1) they lack sufficient capacity to capture long-term dependencies for context modelling, which is critical for SER because human emotions are typically highly context-dependent; and 2) they do not explore the model's dynamic receptive field, whereas learning dynamic receptive fields rather than maximal ones can improve model generalisation ability to unknown data or corpus. Deep learning (DL) is an emerging study subject in machine learning (ML) that has attracted increased attention in recent years [5]. DL techniques for SER have various advantages over previous methods, including the capacity to recognise complicated structures and features without manual extraction, extract low-level features from raw data, and deal with unlabeled data [2], [6]. The primary goal of this paper is to investigate the various speech feature sets and extraction methods, as well as the impact of using different deep neural network (DNN) architectures to detect spoken emotions such as anger, surprise, happiness, sadness, and neutral state in the ShEMO dataset, which is a Farsi/Persian spoken speech dataset.

II. RELATED WORK

This section comprehensively introduces Transformer uses in various disciplines as well as associated research on structure-based paralinguistic speech modelling.

Transformer in Language Processing and Computer Vision

First, the original Transformer is intended to handle machine translation jobs in the realm of natural language processing [26]. Transformer is an excellent sequence learning model for modelling long-term relationships, and because it is entirely based on the attention mechanism, it avoids recursion and convolution and computes hidden representations in parallel. The raw text signal is generally transformed into a word embedding sequence initially by the word and position embedding layers. The output is then sent to a stack of Transformer encoders to generate the final embedding, which is then followed by numerous Transformer decoders or a task-specific classifier. The Transformer has been used for a variety of NLP tasks such as question answering [10], named entity recognition [11], natural language inference [12], semantic textual similarity [13], and document categorization.

Transformer in Paralinguistic Speech Processing

There has also been a great deal of interest in using the standard Transformer in the speech domain. In general, the raw speech stream is divided into many overlapping frames. Then, from each frame, the spectral or deep learning-based features are retrieved and used as input for the Transformer [14-16], [6]. [5] investigated the use of stacked multiple Transformer layers to improve the information collected for voice emotion recognition. Researchers in [6] followed the Swin Transformer [8] construction and sliced the spectrogram into distinct patch tokens for sound classification and detection. Although these works indicate its usefulness, they primarily use Transformer directly, ignoring speech and task factors. To address this issue, [14] used a sparse Transformer to implement efficient emotion recognition. to focus more on the emotion-related information. [17] investigated and applied an auditory saliency method in a Transformer to alter feature embeddings. Transformer has also been used to diagnose Alzheimer's disease (AD) [18], [19], and to classify depression [20]. Ilias et al. [18], for example, used a pretrained vision Transformer [7] to extract audio features and achieved excellent results for dementia detection. Later, Zhu et al. [9] attempted to integrate semantic and non-semantic speech information effectively. Both [18, 19] advocated for the use of pretrained models. [10] employed a Transformer- based network to extract long-term temporal context information for depression estimate. Various self-supervised voice representation learning techniques based on Transformer have also been presented, notably wav2vec [1], wav2vec 2.0 [2] and [3] HuBERT. Several studies based on pretrained self-supervised models have yielded encouraging results in the literature [13], [14], [19], [9], [4]-[17]. Monica et al. [33]
typically refined the pretrained HuBERT model for AD detection and reached competitive performance. Transformer is widely employed in voice recognition [8]-[10] in addition to paralinguistic tasks. Wang et al. [9] investigated the potential of Transformer-based acoustic models on hybrid speech recognition and achieved considerable word error rate improvements over conventional baselines. Gulati et al. [20] suggested a unique convolution-augmented.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

Temporal-aware Bi- direction Multi-scale Network

We offer TIM-Net, a unique temporal emotional modelling approach that learns long- term emotional relationships from forward and backward directions and captures multi-scale aspects at the frame level. The detailed network architecture of TIM-Net is shown in Fig. 1. The TIMNet comprises of n Temporal-Aware Blocks (TABs) in both forward and backward directions with various temporal receptive fields for learning multiscale representations with long-range dependencies. Following that, we go through each component in further depth.Block that is aware of time. The TAB is designed to capture dependencies between frames and automatically choose affective frames, serving as a key unit of TIM-Net. T signifies a TAB, each of which consists of two sub-blocks and a sigmoid function () to train temporal attention maps A, resulting in the temporal-aware feature F by producing the input element by element and A. Each identical subblock of the j-th TAB Tj begins by inserting a DC Conv with the exponentially growing dilated rate 2 j1 and causal constraint. The dilated convolution enlarges and refines the receptive field, while the causal constraint prevents future information from being leaked to the past. Following the DC Conv, there is a batch normalisation, a ReLU function, and a spatial dropout.

III. DEEP LEARNING MODELS

Deep learning has been used efficiently by researchers in recent years due to its multi-layer structure and efficient results in a range of domains, including emotion recognition in speech [1]-[3], [6].DNNs are built with feed-forward structures that include one or more hidden layers between input and output. CNNs are another sort of deep learning approach that is only utilised for classification with forward-looking architecture. CNNs are frequently used to detect patterns and improve classification. RNNs are a type of sequential information neural network in which the outputs and inputs are interdependent, and this reliance is often useful in forecasting the next state of the input. RNNs, like CNNs, require memory to retain generic information gleaned from a sequential deep learning modelling process. procedure, and they usually only work well for a few generations. The biggest issue affecting RNN's overall performance is its susceptibility to gradient disappearance, which results in forgetting the initial input. LSTM is used to construct a block between frequent connections to prevent this. Bidirectional-LSTM networks can also be employed [2, 5]. In recent years, there has been a lot of interest in the merging of CNN and LSTM networks. In the SER task, it is anticipated that CNN extracts specific patterns in the utterance that carry emotional information, while LSTM focuses on the temporal behaviour of the utterance [6], [10]. As a result, the CNN-LSTM architecture can be useful in categorising LLD characteristics. Attention processes in neural networks have demonstrated extensive success in a variety of activities, including speech recognition, machine translation, natural language understanding, and question answering. The basic goal of the attention mechanism is to ignore the rest and concentrate on a few closely connected components. This method comes in several forms (global vs. local, soft vs. hard), but its primary function is to support various LSTM models, such as encoder-decoder architecture (for instance, in machine translation), by preventing the usage of a fixed context vector as the decoder's sole output. This particular hidden LSTM layer carries all of the data that the LSTM encoder has extracted. In the traditional structure, all data is compressed into a context vector, which can be a bottleneck, and all encoder's hidden intermediary layers are disregarded. The LSTM or dense decoder are the next layers to receive this vector. Only this kind of summary provided by the encoder is used in subsequent steps, therefore the performance of the model can be hindered by lengthening the time sequence under analysis.

IV. METHODOLOGY

Three major elements and four stages make up the suggested framework. For structure- based speech unit learning, the unit encoder and word encoder are employed, and for the aggregation of structure-based speech units, a merger block is used. We begin by outlining the standards for model design. Then, we go into further detail about the suggested Speech Former++.

Guiding Principles of Model Design

Our architecture is designed around the statistical duration of the speech unit. As a result, we begin by estimating the lengths of phones and words on the corpora utilised in this work using the P2FA [68] toolset. Because the distribution of unit duration is similar across corpora, we show the statistical results obtained by integrating all audio recordings

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

from four corpora in Fig. 3. Because more than 80% of phones range between 50 and 200 ms, we estimate the shortest and longest durations of phones to be 50 and 200 ms, respectively. Similarly, about 90% of words have durations ranging from 250 to 1000 ms, which we consider to be the shortest and longest.

Table 1: *The data distribution of the IEMOCAP dataset.*

Session	1	2	3	4	5
No.utterance	1085	1023	1151	1031	1241
No.dialogue	28	30	32	30	31

Structure-Based Speech Unit Learning

We initially extract the acoustic representations of a speech signal, x1 R T1d1, where T1 is the number of frames and d1 is the dimension of each frame embedding. We use a unit encoder with window Tw1 to learn the frame- grained features in x1 to record information about consecutive frames in the frame stage. In particular, the frame grained input feature x1 is divided into T1 overlapping segments. where subscript i represents the various steps in Fig 2 (for example, i = 1 for the frame stage, i = 2 for the phone stage, i = 3 for the word stage, and i = 4 for the utterance stage); Overlap Seg() illustrates overlapping segmentation, and j [1, Ti]; xi [a: b] R (ba)di is made up of xi's a-th to b-th tokens. Because it is now in the frame stage, the subscript i is equal to 1. When the segment is out of range (e.g., a 0 or b > Ti), zero padding is used. Tw1 is set to the number of tokens that may be contained during 50 ms (the shortest phone time) of input x1. As a result, the interactions of neighbouring frames

Table 2: *Experimental results of WA(%) for two systems under different training settings.*

	TS_1234	TS_123	TS_134	TS_234	TS_23
Cmp	81.06	80.82	79.85	78.89	77.60
Our	**81.14**	**82.68**	**82.27**	**82.43**	**81.39**
Δ	+0.08	+1.86	+2.42	+3.54	+3.79

Structure-Based Speech Unit Aggregation

Inspired by the property of hierarchy We propose a merging block to generate the relevant characteristics under the statistical instruction of voice signals that may be gradually classified into frames, phones, and words durations of the speech units. Merging blocks are employed between each level, as seen in Fig. 2. The acoustic input of frame stage x1 initially represents the attributes of each frame from the original voice signal. We use average pooling over the output of the frame stage x 1 with a merging scale M1 of 50 to supply the phone-grained input to the phone stage. ms (the shortest phone time). The phone grained feature x2 is then created using a linear projection and layer normalisation. Each token in x2 represents the information of a sub phoneme since the information contained every 50 ms is consolidated into a token in x2. Similarly, when attempting to generate the word-grained input x3 for the word stage, the merging scale M2 is set to 250 ms (the shortest duration of words), resulting in each token in x3 representing a sub word. Finally, the final merging block is applied to the word stage x 3 output while merging scale M3 is set to 1000 ms (the longest duration of words) to roughly simulate the amount of words in the utterance sample.

Table 3: *The performance of state-of-the-art approaches and the proposed approach on the IEMOCAP database.*

Approaches	WA (%)
Abdelwahab et al. (2018) [19]	56.68
Li et al. (2019) [18]	58.62
Rozgić et al. (2012) [29]	67.40
Jin et al. (2015) [30]	69.20
Poria et al. (2017) [11]	74.31
Li et al. (2018) [31]	74.80
Hazarika et al. (2018) [17]	77.62
Li et al. (2019) [32]	79.20
Proposed method	**82.68**

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

V. IMPLEMENTATION DETAILS

Acoustic characteristics: Motivated by the success of self-supervised learning models in a variety of speech tasks, we employ the pretrained HuBERT-large [53] model to extract acoustic characteristics.

The duration of each frame analysed by HuBERT is 25 ms, and the hop length used when yielding the time between overlapping frames is 20 milliseconds. The time difference between consecutive frames is 5 milliseconds. For each utterance sample, 1024-dimensional frame-grained characteristics are extracted. It was recently revealed that the middle layer output has the greatest pronunciation-related properties [76]. As a result, in HuBERT, we use the output from the 12-th layer of the 24-layer Transformer encoder. Unless specified differently, the pretrained self-supervised models are only used to extract acoustic information traits and will be excluded from the training method. The maximum sequence lengths are 326, 224, 328, and 426 because 80% of samples in each dataset are shorter than the appropriate set sequence lengths for IEMOCAP, MELD, Pitt, and DAIC-WOZ, respectively. We also present Speech Former++ results with hand-crafted features, such as 80-dimensional log-mel filter bank coefficients (FBANK). Hubert features are used in Speech Former++ unless otherwise specified.

TABLE I
PERFORMANCE AND COMPUTATIONAL EFFICIENCY OF TRANSFORMER AND SPEECHFORMER++ USING HUBERT FEATURES ON IEMOCAP. GAIN INDICATES THE RELATIVE IMPROVEMENT (+) OR REDUCTION (-)

Method	Params	FLOPs	WA	UA	WF1
Transformer	63.64M	23.12G	0.685	0.701	0.692
SpeechFormer++	66.79M	6.55G	**0.705**	**0.715**	**0.707**
Gain	+4.95%	-71.67%	+2.92%	+2.00%	+2.17%

Training Details

We train Speech Former++ from start to finish on an Nvidia GeForce RTX 2080 Ti GPU. For IEMOCAP, MELD, Pitt, and DAIC-WOZ, the total number of training epochs is set to 120, 120, 80, and 60, respectively, and the initial learning rates are set to 0.0005, 0.0005, 0.001, and 0.0001, respectively. Cosine annealing reduces the learning rate to 1% of the original. The batch size has been set to 32. SGD updates the model with momentum 0.9. In MAS, the number of attention heads is fixed to eight. To keep things simple, the dimensions of xi and zi, i 1, 2, 3, 4, are all set to 1024. Unless otherwise specified, the number of layers in the frame stage N1, phone stage N2, and word stage N3 is 2, 2, and 4, respectively, and the number of Transformer encoders in the utterance stage N4 is 4. As a result, Speech Former++ has a total of 12 levels.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON IEMOCAP. ALL SYSTEMS APPLY AUDIO AS INPUT FOR A FAIR AND DIRECT COMPARISON. H/C=HAND-CRAFTED, W2V2=WAV2VEC 2.0

Method	Features	Year	WA	UA	WF1
STC [14]	H/C	2021	0.613	0.604	0.617
† ISNet [15]	H/C	2022	0.704	0.650	-
LSTM-GIN [17]	H/C	2021	0.647	0.655	-
SUPERB [77]	w2v2	2021	0.656	-	-
SUPERB [77]	HuBERT	2021	0.676	-	-
CA-MSER [54]	H/C + w2v2	2022	0.698	0.711	-
SpeechFormer++	H/C	2022	0.645	0.658	0.649
SpeechFormer++	HuBERT	2022	**0.705**	**0.715**	**0.707**

† Speaker information is used.

VI. CONCLUSION

We propose a context-dependent domain adversarial neural network for multimodal emotion recognition in this paper. We run trials on the IEMOCAP database to see how effective our proposed strategy is. The experimental results show that our strategy allows the model to focus on emotion-related information while ignoring the difference between speaker identities. This is why, when compared to the fully supervised learning technique, we gain greater performance on unseen speakers. Meanwhile, our approach can correctly use unlabeled materials and generate good results in low-resource conditions. Furthermore, we show that using contextual and multimodal information in DANN can increase emotion recognition performance. Because of the benefits listed above, this unique approach outperforms state-of-the-art methodologies for emotion recognition. Without employing explicit linguistic information, we achieve state-of-the-art valence recognition performance on MSP-Podcast of.638 and credit this extraordinary result to implicit linguistic

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

information learned through fine-tuning of the self-attention layers. We make available to the community our highest performing model (w2v2- L-robust-12) [26]. Transformer topologies are more resistant to tiny perturbations, perform well in (biological sex) groups if not individuals, and generalise across domains. Our findings show that a new era in speech emotion detection is dawning: the era of pre-trained, transformer-based foundation models, which can ultimately lead to the coveted integration of the two primary information streams of spoken language, linguistics, and paralinguistics

REFERENCES

[1] B. Moore, L. Tyler, and W. Marslen-Wilson, "Introduction. The perception of speech: from sound to meaning," Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 363, no. 1493, pp. 917–921, Mar. 2008.
[2] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 1995, pp. 660–663.
[3] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," IEEE Transactions on Signal Processing, vol. 46, no. 4, pp. 886–902, 1998.
[4] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings, vol. 1, 2003, pp. I–401.
[5] J. Cichosz and K. Slot, "Emotion recognition in speech signal using emotion-extracting binary decision trees," Proceedings of affective computing and intelligent interaction, 2007.
[6] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to- end dialogue systems with commonsense knowledge," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 4970–4977.
[7] L. Chen, C. Chang, C. Zhang, H. Luan, J. Luo, G. Guo, X. Yang, and Y. Liu, "L2 learners' emotion production in video dubbing practices," in IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2019, pp. 7430–7434.
[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in Proceedings of Ninth European Conference on Speech Communication and Technology, 2005.
[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, p. 335, 2008.
[10] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in IEEE International Conference on Acoustics, Speech and Signal Processing, 1997, pp. 1039– 1042.
[11] Zixuan Peng, Yu Lu, Shengfeng Pan, and Yunfeng Liu, "Efficient speech emotion recognition using multi-scale CNN and attention," in ICASSP 2021, Toronto, ON, Canada, June 6- 11, 2021. 2021, pp. 3020–3024, IEEE.
[12] Linhui Sun, Sheng Fu, and Fu Wang, "Decision tree SVM model with Fisher feature selection for speech emotion recognition," EURASIP J. Audio Speech Music. Process., vol. 2019, pp. 2, 2019.
[13] Luefeng Chen, Wanjuan Su, Yu Feng, Min Wu, Jinhua She, and Kaoru Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," Inf. Sci., vol. 509, pp. 150–163, 2020.
[14] Jiaxin Ye, Xin-Cheng Wen, Xuan-Ze Wang, Yong Xu, Yan Luo, Chang-Li Wu, Li-Yan Chen, and Kunhong Liu, "GMTCNet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition," Speech Commun., vol. 145, pp. 21–35, 2022.
[15] J Ancilin and A Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," Applied Acoustics, vol. 179, pp. 108046, 2021.
[16] Arya Aftab, Alireza Morsali, Shahrokh Ghaemmaghami, et al., "LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition," in ICASSP 2022, Virtual and Singapore, 23-27 May 2022. 2022, pp. 6912–6916, IEEE.
[17] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria: ISCA, 2019, pp. 2823–2827.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

[18] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.
[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria (virtual), 2020, pp. 1597– 1607.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008.