QUATERNION ANTI-TRANSFER LEARNING FOR SPEECH EMOTION RECOGNITION

Eric Guizzo, Tillman Weyde, Giacomo Tarroni

Department of Computer Science
City, University of London
London EC1V 0HB, UK
{eric.guizzo, t.e.weyde,
giacomo.tarroni}@city.ac.uk

Danilo Comminiello

Department of Information Engineering
Electronics and Telecommunications (DIET)
Sapienza University of Rome
Rome 00184, Italy
danilo.comminiello@uniroma1.it

ABSTRACT

This study explores the benefits of anti-transfer learning with quaternion neural networks for robust, effective, and efficient speech emotion recognition. Anti-transfer learning selectively promotes task invariance through the introduction of a deep feature loss at training time. It has been shown to improve the performance of speech emotion recognition models by encouraging the independence of emotion predictions from specific uttered words and characteristics of the speaker's voice. However, the improved accuracy comes at a cost of increased computation time and memory requirements. In order to reduce the resource demand of anti-transfer, we propose to exploit quaternion-valued processing. We design, implement, and evaluate the use of quaternion anti-transfer learning on the basis of the VGG16 architecture and quaternion embeddings on multiple datasets for different speech emotion recognition task setups. The effectiveness of this approach depends on the layer where it is applied, with early layers offering a good compromise between performance gain and resource requirements. Our results show that anti-transfer in the quaternion domain can enhance generalisation while reducing the model's demand for computation and memory.

1. INTRODUCTION

Research on automatic speech recognition reached near-human performance in recent years, making it possible to confidently identify which words are uttered even in audio signals with non-optimal recording quality and for an extensive amount of languages [1]. Nevertheless, without the understanding of the speaker's non-verbal behavior this information is incomplete, and recognition of the emotional intention helps provide a fuller account of the spoken communication [2]. Emotion recognition can be ambiguous even for humans, since it strongly depends, among other factors, on the context, on the speaking style of a person, and on their cultural background [3]. Despite recent success with neural networks, speech emotion recognition (SER) is still challenging due to the variability of emotional expression, especially when generalization to unseen speakers and contexts is required [4, 5].

Several different solutions have been proposed to make SER models robust to unseen speakers and context. One approach is to design signal representations for robustness. Among others, [6] propose a hard-coded audio feature, the *ratio of a spectral flatness measure to a spectral center*, to help models generalize to new speakers. [7] propose a Fourier-based *voice quality* feature, which, alongside its derivatives, helps generalization to unseen speakers. [8] propose feature extraction based on wavelet packet coefficient. Another approach is to use machine learning strategies to improve generalisation. [9] uses an ensemble of acoustic and linguistic models merged through late fusion. [10] propose a cascade schema that progresses from macro-categories of emotions gradually towards the discrimination of more specific emotions. [11] proposes to fine-tune several SER classifiers for specific speakers and to select the classifier to use with a speaker recognition system. [12] applies a unsupervised multi-source domain adaptation strategy to learn emotion features independent from the speaker identity. [13] achieves a similar disentangled emotion representation through adversarial training based on a gradient reversal technique to remove speaker information.

This work is based on anti-transfer learning (AT), proposed in [14], which is a supervised machine learning approach to selectively promote task invariance in neural networks through a deep feature loss, the *anti-transfer loss* (ATL). Minimizing ATL leads to the divergence of the features developed in a specific layer of a CNN from the features of a network with identical structure but pre-trained on an orthogonal task, i.e., a task that should not influence the target task. [14] demonstrates that AT can improve the performance of models in different domains and tasks. In particular, it can improve the invariance of SER models to the uttered words in the training data and to characteristics of the speaker's voice that can bias the result, especially when using datasets with few actors and predefined sentences. However, the improved accuracy comes at a cost of increased computation time for training, on average 2.8 times longer per epoch than without AT, and with an additional memory demand of up to 100% [14].

In order to reduce the resource demand of AT, we propose to use quaternion information processing, which is increasingly popular strategy to significantly reduce model parameters while enhancing or maintaining the model's performance [15, 16]. Quaternion information processing allows to exploit inter-channel correlations in neural networks by using the Hamilton product instead of the dot product, which reduces trainable parameters by up to 75% compared to the real-valued counterparts [16, 17], as further explained in the next section. In particular, we apply anti-transfer learning to quaternion-valued CNNs (QCNNs) and quaternion embeddings generated with RH-Emo, as described in [18]. This technique extracts quaternion embeddings enriched with emotion-related information, that enable performing SER tasks with quaternion neural network. Therefore, we develop AT for RH-Emo QCNNs to maintain the benefits of AT while reducing the model's demand for memory and computation time.

Eric Guizzo has been supported by a doctoral study grant from City, University of London. The work of D. Comminiello has been partly supported by Sapienza University grant RG11916B88E1942F.

979-8-3503-2372-6/23/$31.00 ©2023 IEEE

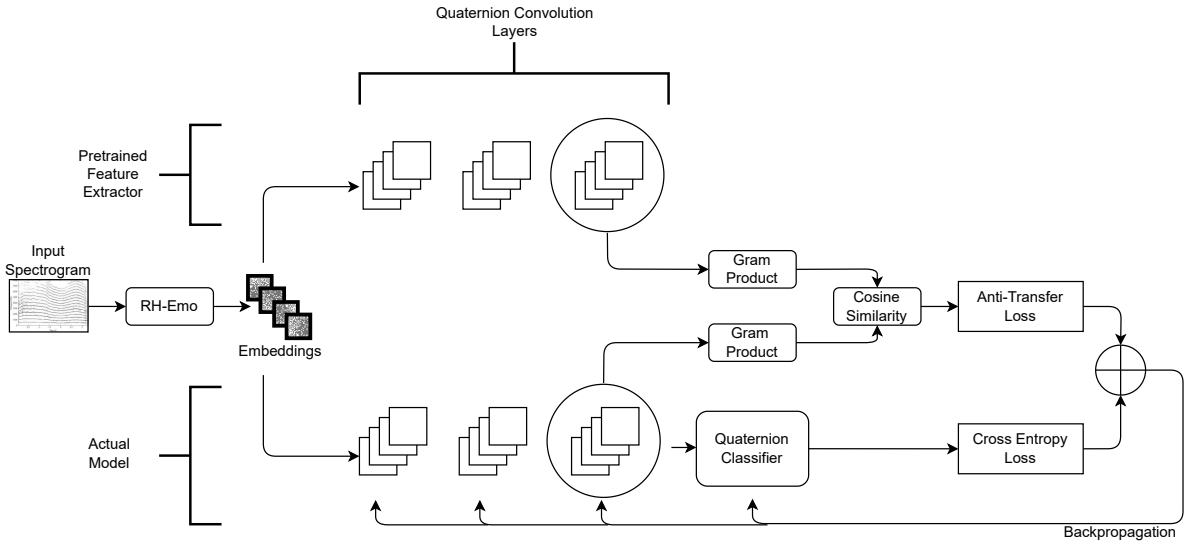

Figure 1: Block diagram of a QCNN network with RH-Emo feature extractor and *anti-transfer learning* applied to SER classification.

2. METHOD

As introduced in [14], the anti-transfer loss $\mathcal{L}_{\text{AT}}$ is defined as the squared cosine similarity of the vectorized Gram matrices G_t and G_p, multiplied by a scalar weighting coefficient β. G_t and G_p computed by the Gram product of the feature maps of a currently trained layer (G_t) with a corresponding layer (G_p), that has been pre-trained on an orthogonal task:

$$\mathcal{L}_{\text{AT}} = \beta \left(\frac{G_p \cdot G_t}{||G_p||\,||G_t||} \right)^2. \tag{1}$$

The $\mathcal{L}_{\text{AT}}$ term is added to the loss function during training. The total objective function L_{TOT} that we minimize per datapoint is:

$$L_{TOT} = -\sum_{i=1}^{n} t_i \log(p_i) + \sum_{s \in S_{AT}} L_{ATs} \tag{2}$$

where n is the number of classes, t_i is 1 if i is the true class and 0 otherwise, p_i is the predicted probability of class i, S_{AT} is the set of convolution layers where anti-transfer is computed, and L_{ATs} is the anti-transfer loss computed for convolution layer s. This loss function is applied in the training of the target model, while for the pre-trained feature extractors, L_{ATs} is not used.

This approach has been applied only to real-valued CNNs, yet, and the objective of this research is to apply the anti-transfer learning principle to quaternion-valued CNNs.

We left the original anti-transfer loss definition unchanged, as it is already compatible with feature maps that are generated by quaternion-valued convolution layers for the following reasons. In the first instance, quaternion convolution layers have the same output shape as corresponding real-valued convolution layers. However, in quaternion convolution layers, kernels are aggregated in groups of 4, where each group uses 1/4 of the sub-matrices (channels) to build its kernels. E.g., a real-valued convolution layer that has 4 input and 4 output channels contains 4 kernels shaped as $4\,channels \times time\,dim \times features\,dim$, i.e., it contains a total of 16 sub-matrices shaped as $time\,dim \times features\,dim$. An equivalent quaternion-valued convolution layer contains only 4 different sub-matrices in total, all kernels are made with the same sub-matrices, but in each kernel, they have a different ordering along the channel dimension [16]. Despite their quadral organization, the output matrices of a quaternion convolution layer can be treated as real-valued matrices, because each channel contains unique information.

We use the Gram matrix in order to compare all possible channel combinations in the ATL, limiting the impact of possible channel permutations. For this, we calculate the Gram matrix on the quaternion components as real numbers. By doing this, we treat correlations within the quaternion axes in the same way as correlations between components of different quaternions which is justified by the unique information contained on each axis.

Figure 1 shows a block diagram of a QCNN with anti-transfer learning. The input magnitude spectrogram is first propagated through a pre-trained RH-Emo network, that generates quaternion embeddings to feed into both the pre-trained feature extractor and the network being currently trained. The anti-transfer loss is then computed exactly as for real-valued CNNs, as described above.

3. EVALUATION

We evaluate the properties of anti-transfer learning applied to quaternion-valued CNNs for SER tasks. Our aim is to compare AT performed in the real and quaternion domains to each other and to the same real and quaternion networks without any AT applied.

3.1. Experimental Setup

In this study, we used 3 popular SER datasets: the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [19], the Ryerson Audio Visual Database of Emotional Speech and Song [20] (RAVDESS) and the Toronto Emotional Speech Set [21] (TESS).

We chose IEMOCAP because it is among the most frequently used SER datasets and also contains improvised material. On the other hand, the others include systematic variation (same words repeated with different emotions and from different speakers), which makes them suitable for studying the specific benefits of AT.

We apply 4 preprocessing stages, as in [18]: we cut signals into non-overlapping frames of 4-seconds length, we compute the STFT using 16ms sliding windows with 50% overlap, we discard the phase information and we zero-pad all matrices to a shape of 512 x 128. In addition to this, we normalize the magnitudes to range 0 to 1 for the quaternion-valued networks (as required by RH-Emo) and to 0 mean and unity standard deviation for real-valued networks, which helps convergence.

For the quaternion networks, we keep the same architecture as for the real-valued ones and use equivalent quaternion-valued convolution and fully connected layers instead of the real-valued operations. There is an exception in the final layer of the networks, which are real-valued also in the QCNNs. For the real-valued networks, we use the magnitudes-only spectra as input, while for the quaternion networks we use a trainable RH-Emo network to generate embeddings. RH-Emo is pre-trained on a dual target of predicting IEMOCAP emotion labels and auto-encoding.

We use the standard Pytorch implementation of the VGG16 network [22] from the torchvision library[1]. In the quternion networks, we remove the adaptive average pooling layer between the convolutional feature extractor and the dense classifier section. This layer serves to reduce the feature extractor output shape to a fixed input shape of the classifier. When the RH-emo embeddings are used as input, however, the feature map is smaller than the classifier input. Removing the adaptive pooling layer and reducing the classifier input thus prevents the generation of redundant output in the adaptive layer and reduces the number of network parameters.

For all experiments, we used the ADAM optimizer, a learning rate of 0.00001, and a batch size of 20 samples. We apply early stopping with a patience of 20 epochs on the validation loss and we split the training, validation, and test sub-sets with approximately 70%, 20%, and 10% of the data, respectively.

We configured the experimental setup in order to show the difference between using AT on real and on quaternion-valued CNNs with embedding input, as well as the difference between using AT or not for the same network architectures. To this extent, we have paid attention to performing all experiments with the exact same configuration in order to isolate the features we want to highlight. This makes our results non-comparable with published state-of-the-art results for the same datasets, as those are obtained with more complex setups that are different for each dataset [23, 24, 25] and recreating them would make it more difficult to isolate and understand the properties of our approach.

We perform AT with two orthogonal tasks: Word Recognition and Speaker Recognition, as in the SER experiments of [14]. We expect that invariance to the spoken words can improve the model's performance in cases where text and speech emotion do not align (such as sarcasm or irony), while invariance to the speaker would help disentangling predictions from specific speech styles and thus improve generalization to unseen speakers. For the first orthogonal task, we pre-trained the network on the Librispeech [26] dataset, where we extracted single-word segments of 1000 different classes (i.e. words).

[1] https://github.com/pytorch/vision/blob/main/torchvision/models/vgg.py

Table 1: Results for different combinations of dataset, network type, and anti-transfer type. **AT L.** shows the layer where AT is applied. **Rnd-s** and **Act-s** are accuracy with random and actor split

Arch.	Method	AT Type	Params	AT L.	Rnd-s	Act-s
IEMOCAP	Real	/	1.6×10^8	/	62.87	58.29
	Real-AT	Single S	1.6×10^8	11,10	72.86	59.27
	Real-AT	Single W	1.6×10^8	1,11	70.09	59.35
	Real-AT	Dual SW	1.6×10^8	11,9	73.04	**60.41**
	Real-AT	Dual WS	1.6×10^8	8,8	71.07	59.67
	Quat	/	1×10^7	/	71.1	57.8
	Quat-AT	Single S	1×10^7	11,12	72.23	58.37
	Quat-AT	Single W	1×10^7	2,13	**73.57**	57.97
	Quat-AT	Dual SW	1×10^7	5,12	71.25	58.46
	Quat-AT	Dual WS	1×10^7	4,9	72.14	58.62
RAVDESS	Real	/	1.6×10^8	/	41.06	45.0
	Real-AT	Single S	1.6×10^8	2,9	49.56	46.67
	Real-AT	Single W	1.6×10^8	9,1	42.89	41.67
	Real-AT	Dual SW	1.6×10^8	2,2	45.56	40.62
	Real-AT	Dual WS	1.6×10^8	1,9	46.22	43.75
	Quat	/	1×10^7	/	49.85	36.67
	Quat-AT	Single S	1×10^7	2,2	57.11	51.67
	Quat-AT	Single W	1×10^7	12,13	55.11	48.33
	Quat-AT	Dual SW	1×10^7	3,2	57.78	**55.0**
	Quat-AT	Dual WS	1×10^7	11,2	**60.44**	46.67
TESS	Real	/	1.6×10^8	/	97.62	14.29
	Real-AT	Single S	1.6×10^8	3,4	99.64	13.29
	Real-AT	Single W	1.6×10^8	3,4	99.64	**26.43**
	Real-AT	Dual SW	1.6×10^8	3,2	**100.0**	21.0
	Real-AT	Dual WS	1.6×10^8	3,6	**100.0**	25.57
	Quat	/	1×10^7	/	97.62	13.86
	Quat-AT	Single S	1×10^7	1,11	97.86	19.43
	Quat-AT	Single W	1×10^7	11,4	98.57	18.29
	Quat-AT	Dual SW	1×10^7	2,11	98.93	19.0
	Quat-AT	Dual WS	1×10^7	9,3	97.86	18.29

3.2. Experimental Results

Table 1 shows our performance results for IEMOCAP, RAVDESS and TESS. We compare real and quaternion-valued networks without AT and with AT applied in different recognition tasks: *Single S*:Single AT, speaker recognition, *Single W*: Single AT, word recognition, *Dual SW*: Dual AT first speaker, then word recognition, *Dual WS* Dual AT first word, then speaker recognition.

In Dual AT we train a model with anti-transfer for one orthogonal task and using the result of that training to initialize the weights of a new model, which is then trained with anti-transfer on the second orthogonal task [14]. We use the same AT layer on both orthogonal tasks when we perform dual AT. The table exposes the test accuracy obtained by splitting the training, validation and test set randomly (*Rnd-s* column) and actor-wise ((*Act-s* column). While in the first split setting the samples recorded from all actors appear in all sets, in the latter setting the models are tested on speakers unseen during the training stage, and therefore better show the model's generalization capabilities. In each case, we select the AT layer that gives the best validation accuracy (even though we report the test accuracy). Column "AT L." lists the layers that provide the reported Rnd-s and Act-s results, respectively.

3.2.1. *Effect of anti-transfer*

In general, the experimental results we obtained demonstrate that AT can be successfully applied to quaternion-valued neural networks, even though the average accuracy performance improvement is higher for real-valued ones. This further supports the foundation of AT, showing that it is not restricted to the real-valued networks presented in [14]. AT provides a consistent improvement both for real and quaternion-valued networks. However, among the cases considered in this study, there is not a single best approach overall. For all datasets, the best test accuracy result is provided by a network where AT is applied. AT provides an average improvement of 3.66 percentage points (pp) for real-valued networks and of 2.6 pp for quaternion-valued networks. Single AT achieves an average improvement of 3.12 pp, whereas the improvement for dual AT is of 3.15 pp. In the random split configuration the average improvement is 3.82 pp, instead on the actor-wise split setup it is limited to 2.45 pp. AT has a different impact on each dataset, providing an average performance boost of 3.01 pp, 2.62 pp and 3.77 pp for IEMOCAP, RAVDESS and TESS, respectively. Furthermore, the 2 different orthogonal tasks used to pre-train the AT classifiers also have different impact on the models' performance: speaker recognition boosts the average test accuracy by 2.93 pp and word recognition by 3.33 pp. In this setup, it could be expected that the RH-Emo pre-training on IEMOCAP gives an advantage in the RAVDESS and TESS case, because additional data is used. At least for TESS, this does not appear in the data.

We have performed further experiments where we pretrain on IEMOCAP and apply transfer learning by weight initialization to our RAVDESS and TESS baselines and found that the impact of RH-Emo pretraining does not change the overall picture: quaternion AT provides similar accuracy to real-valued networks and significantly improves efficiency.

3.2.2. *Per-layer AT performance*

Figure 2 shows the average test accuracy using AT at different layers in our networks with different setups: random vs actor split, speaker vs word recognition AT pre-training, and real vs quaternion-valued networks. The 5 colored lines indicate the combination of orthogonal task and train/test split types adopted. Each point in these lines shows the average test accuracy obtained for a specific AT layer among all corpora and the green dotted line draws the mean value among all visible lines.

The Figure shows that the impact of AT is different for different setups and convolution layers. AT does not provide a performance improvement in all cases, confirming the importance of AT layer selection (see [14]). Nevertheless, real and quaternion networks exhibit a similar trend with an drop in performance for AT in the middle layers. On real-valued networks the random split configuration shows the highest and more consistent improvement, whereas for quaternion networks the actor-split setup, which is more relevant in practive, provides the greatest improvement. These improvements do not always indicate that a quaternion network gives the best overall performance, because of the different baselines and the fact that the model selection is based on the validation set while Figure 2 shows test set results. Taking into account that AT in the earlier layers need less resources, the results show that the early layers (layer 2 in particular) offer a good trade-off between performance gain and resource requirement for QCNNs.

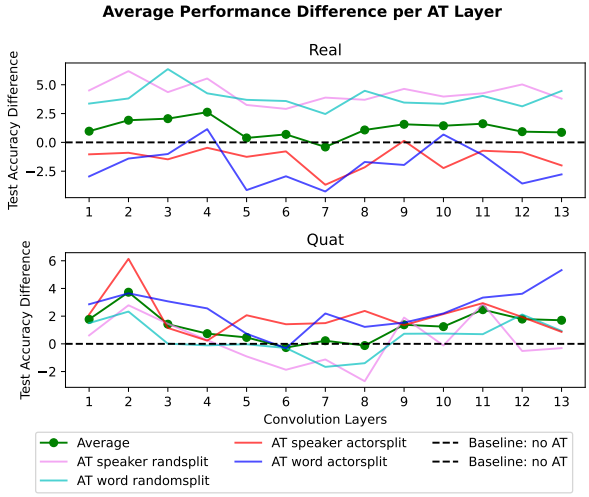

Figure 2: Test accuracy for different AT layers in different setups.

3.2.3. *Resource demand*

We see a significant reduction in both memory requirements and training times when using quaternion-valued networks. Quaternion-valued VGG16 networks reduce training times on average by 80.5% and memory demand by 53.7%, compared to real-valued networks.

The deeper is the AT layer, the higher are the average training time and memory demand, both for real and quaternion-valued networks. This is exacerbated by VGG16's structure with an increasing number of channels, resulting in larger Gram matrices.

We observe that the quaternion setup without AT has a worse generalisation to new speakers (Act-s) than the real-valued variant. However, applying AT always more than compensates for that. Testing for the optimal layer can be an expensive strategy at training time. However, QCNN and RH-Emo with AT applied to early layers appears to be a good heuristic, as it provides a good trade-off between reducing computational cost at both train and run time with performance improvement.

4. CONCLUSIONS

In this paper, we propose a novel approach to exploit the benefits of QCNNs with RH-Emo embeddings to reduce the impact of anti-transfer learning on resource demand. This method retains most anti-transfer benefits, while requiring a fraction of memory and computation time. The experimental results demonstrate that AT can be successfully applied to RH-Emo QCNNs, obtaining a significant improvement in the test accuracy performance for SER tasks. The combination of AT and QCNNs fed with RH-Emo embeddings saves a significant amount of resources: up to 80.5% of the training time and up to 53.7% of memory, making this approach a relevant option to exploit the disentanglement properties of AT in scenarios with restricted resources. Further work will focus on extending the proposed method to other tasks and other network architectures.

5. REFERENCES

[1] Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom, "Automatic speech recognition: A survey," *Multimedia Tools and Applications*, vol. 80, pp. 9411–9457, 2021.

[2] Albert Mehrabian and Morton Wiener, "Decoding of inconsistent communications.," *Journal of personality and social psychology*, vol. 6, no. 1, pp. 109, 1967.

[3] Klaus R Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[4] Dias Issa, M Fatih Demirci, and Adnan Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. and Control*, vol. 59, pp. 1–11, 2020.

[5] Mehmet Berkehan Akçay and Kaya Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[6] Eun Ho Kim, Kyung Hak Hyun, Soo Hyun Kim, and Yoon Keun Kwak, "Improved emotion recognition with a novel speaker-independent feature," *IEEE/ASME transactions on mechatronics*, vol. 14, no. 3, pp. 317–325, 2009.

[7] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li, "Speech emotion recognition using Fourier parameters," *IEEE Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[8] Kunxia Wang, Guoxin Su, Li Liu, and Shu Wang, "Wavelet packet analysis for speaker-independent emotion recognition," *Neurocomputing*, vol. 398, pp. 257–264, 2020.

[9] Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Eurospeech*, 2005.

[10] Margarita Kotti and Fabio Paternò, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *International journal of speech technology*, vol. 15, no. 2, pp. 131–150, 2012.

[11] Lyudmila Savchenko and Andrey V Savchenko, "Speaker-aware training of speech emotion classifier with speaker recognition," in *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*. Springer, 2021, pp. 614–625.

[12] Cheng Lu, Yuan Zong, Wenming Zheng, Yang Li, Chuangao Tang, and Björn W Schuller, "Domain invariant feature learning for speaker-independent speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2217–2230, 2022.

[13] Haoqi Li, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7144–7148.

[14] Eric Guizzo, Tillman Weyde, and Giacomo Tarroni, "Anti-transfer learning for task invariance in convolutional neural networks for speech processing," *Neural Networks*, vol. 142, pp. 238–251, 2021.

[15] Yi Tay, Aston Zhang, Luu Anh Tuan, Jinfeng Rao, Shuai Zhang, Shuohang Wang, Jie Fu, and Siu Cheung Hui, "Lightweight and efficient neural natural language processing with quaternion networks," in *Proc. of the 57th Ann. Meeting of the Assoc. for Computat. Linguistics*, 2019, pp. 1494–1503.

[16] Eleonora Grassucci, Aston Zhang, and Danilo Comminiello, "PHNNs: Lightweight neural networks via parameterized hypercomplex convolutions," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2022.

[17] Danilo P Mandic, Cyrus Jahanchahi, and C Cheong Took, "A quaternion gradient operator and its applications," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 47–50, 2010.

[18] Eric Guizzo, Tillman Weyde, Simone Scardapane, and Danilo Comminiello, "Learning speech emotion representations in the quaternion domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1200–1212, 2023.

[19] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] Steven R Livingstone and Frank A Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS One*, vol. 13, no. 5, 2018.

[21] Kate Dupuis and M Kathleen Pichora-Fuller, *Toronto Emotional Speech Set (TESS)*, University of Toronto, Psychology Department, 2010.

[22] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. on Learning Representations (ICLR)*, 2015.

[23] Sofoklis Kakouros, Themos Stafylakis, Ladislav Mosner, and Lukas Burget, "Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing," *arXiv preprint arXiv:2211.01756*, 2022.

[24] Yassamine Lala Bouali, Olfa Ben Ahmed, and Smaine Mazouzi, "Cross-modal learning for audio-visual emotion recognition in acted speech," in *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE, 2022, pp. 1–6.

[25] S Jothimani and K Premalatha, "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons & Fractals*, vol. 162, pp. 112512, 2022.

[26] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.