INSIGHTS OF NEURAL REPRESENTATIONS IN MULTI-BANDED AND MULTI-CHANNEL CONVOLUTIONAL TRANSFORMERS FOR END-TO-END ASR

A. Ollerenshaw[1], *M.A. Jalal*[1], *T. Hain*[1]

[1]Speech and Hearing Research Group, University of Sheffield, UK

ABSTRACT

End-to-End automatic speech recognition (ASR) models aim to learn generalised representations of speech. Popular approaches for End-to-End solutions have involved utilising extremely large amounts of data and large models to improve recognition performance. However, it is not clear if these models are generalising the training data or memorising the data. This paper combines the power of a mixture of experts (MoE) models, which is referred to as multi-band, multi-channel, with a popular model for ASR, the CNN-transformer, to capture longer-term dependencies without increasing the computational complexity of training. The goal is to investigate how the transformer models adapt to these different input representations of the same data. No external language models were used to remove the impact of external language models during inference. Although the proposed multi-band transformer shows performance gain, the main finding of this paper is to show the adaptive memorisation nature of transformers and the neural representations of transformer embedding. Using the statistical correlation index SVCCA, comparative discussion of the neural representations of the proposed model and transformer approach is provided, with key insights into the distinct learned structures.

Index Terms— end-to-end, automatic speech recognition, transformer, interpretability, convolutional neural networks

1. INTRODUCTION

In recent years, the main approaches for automatic speech recognition (ASR) solutions have been hand-crafted deep neural networks (DNNs) combined with Hidden Markov models (HMMs) and End-to-End models, which train all modules jointly in a globally optimised method. The performance of the End-to-End approaches are now level to DNN-HMM models when a sufficient amount of training data is utilised [1]. However, the performance of End-to-End models is still comparatively worse on lower resource tasks or with particularly challenging data.

Particular focus of current developments with attention-based models have involved data augmentation techniques [2], [3], [4] and vastly increasing model depths [5], [6] in an attempt to provide richer neural representations. However, it is not always clear whether these models are generalising or memorising the data and the performance improvements are not attributed to the structures within the actual model architecture. A current state-of-the-art approach [7] utilises the combination of convolutional neural networks (CNNs) and a transformer to provide further improved ASR performance. It is hypothesised that this is due to the ability of CNNs to capture richer local feature representations while the transformer is better able to capture global context. The Linformer model [8] attempted to approximate the information in the attention matrix of the original transformer model. This was done by linearly scaling the attention by projecting the embedding matrix into lower dimension space then computing the inner product. A further attempt from [9] aimed to remove the independence assumptions during modelling to capture long-term context dependencies for End-to-End models. This approach used a “knowledge distillation” technique, where a hierarchical transformer model handles utterance level contextual information and discourse level information independently, while sharing the learned dependencies.

This paper explores the implementation of scalable multi-band CNN models to capture longer-term dependencies, inspired by a mixture of experts (MoE), which has been shown to be effective in NLP [10] [11] and vision domains [12]. This approach aims to retain the model representation capacity while keeping the inference cost constant by applying a subset of parameters to each sample.

Furthermore, building upon previous work from [13], the neural representations of the multi-band model are compared to observe the interaction between the developed structures and the data. SVCCA has been used previously [14] to compare DNN representations. This work aims to provide further insights on the similarity of the learned structures across training and provide a discussion on the distinct representations that occur within convolutional-transformer models and the adaptive memorisation capability of the transformers.

This work was partly supported by Voicebase Inc. at the Voicebase Research Center

2. MODEL ARCHITECTURE

Recently CNNs have been shown to improve ASR model performance when combined with transformer models compared to recurrent based models as they are able to capture local feature information progressively, while the transformer is better able to handle the longer range global context. Variations of this approach, such as [7], combine the convolutions with the self-attention mechanism of the transformer to achieve state-of-the-art results on ASR tasks.

2.1. Multi-Band and Multi-Channel Convolutions

As convolutional networks process the entire spectogram of the audio signal with the same time-frequency resolutions, number of filters, and dimensionality reduction, previous work [15] has shown that higher resolution features can be extracted if the lower frequency bands are processed with high frequency resolution filters and high frequency bands with high time resolution filters. This is due to there being more "voice information" in the lower frequency bands than the higher bands. Furthermore, [16] found that deeper transformer layers dilute audio features, and that the distinction is more profound with spontaneous conversational speech.

The multi-band features f are defined as having N sub-bands, with filterbanks over C channels. The i_{th} filterbank of the j_{th} band of the frame of speech can be described by:

$$f_i^{(j)} = W_{C,i}^T x_C^{(j)} \tag{1}$$

where $W_{C,i}$ is the discrete cosine transform function:

$$W_{C,i} = \sqrt{\frac{2}{C}} \cos\left[(k - 0.5)\frac{i\pi}{C}\right] \tag{2}$$

and where k is the channel energy amplitude.

By modifying the fully-connected convolutional layers with separate filters, features can be extracted at multiple levels of the frequency spectrum. The output layers can then be concatenated together. The proposed architecture is described by Figure 1.

Along a similar methodology, a multi-channel (mchan) approach takes the entire input into parallel convolutional blocks, in an attempt to learn different representations of the same acoustic signal. The representations are then aggregated using mixture of experts in the same method as the multi-band approach. Instead of taking the frequency bands as different streams, as shown by Figure 1, the whole input is taken in multiple streams.

2.2. Encoder-Decoder Transformer

Transformer models are currently the predominant choice for a multitude of domains, such as image recognition and speech recognition. The model published in [17] has especially been utilised for End-to-End speech recognition due to its ability to create a more accessible parallel training method which has allowed End-to-End solutions to make use of larger amounts of data. The main component of the transformer model is the attention module, which measures the similarity of pair-wise positions for a window of the input sequence. The transformer model has an encoder-decoder structure with stacked self-attention and point-wise, fully connected layers. Each of the blocks has a multi-head self-attention layer and feed-forward layer.

The positional encoder takes the filtered output of the convolutional layers to determine the context based on the position of features in the embedding space. The context embedding is then passed to the encoder block, where it is fed through a multi-head attention layer and feed forward layer. The self-attention mechanism aims to attend across a window of the input with reference to the rest of the input. Each attention vector is then finally passed through a feed-forward layer to continue the sequence of encoder blocks. The output of the encoder is then passed to a decoder block where a self-attention layer produces another attention vector over these embedding vectors. These embeddings are then passed to another attention block to determine the relationship between the input and output sequences. Finally, the embedding is then passed through a final feed-forward unit to expand the dimensions into the target output size and normalised, typically with a softmax function.

3. NEURAL INSIGHTS WITH SVCCA INDEX FOR END-TO-END ASR

The task of End-to-End ASR is to identify the acoustic sequence $X = \{x_1, ..., x_T\}$ for time T as the output label sequence $Y = \{y_1, ..., y_N\}$ of length N to map the posterior $p(Y|X)$. Using a statistical correlation method such as singular value decomposition with canonical correlation analysis (SVCCA) [14], two sets of observations can have their correlation relationship measured. For the dataset X and neuron i in layer l, the activation output can be defined as $z_i^l = (z_i^l(x_1), ..., z_i^l(x_T))$. In this case, SVCCA is used to find the two bases w and s, such that when the original matrices are projected onto them, the correlation is maximised:

$$\frac{w^T \sum_{XY} s}{\sqrt{w^T \sum_{XX} w}\sqrt{s^T \sum_{YY} s}} \tag{3}$$

where $\sum_{XX}, \sum_{XY}, \sum_{YY}$ are the covariance and cross-covariance respectively. The projections of the layers l_1 and l_2 are then pruned to the top 99% representative dimensions. The correlation is then calculated by maximising the correlation of the projections of the linear transformations of the layers l_1', l_2':

$$\rho = \frac{\langle w^T l_1', s^T l_2' \rangle}{||w^T l_1'||\ ||s^T l_2'||} \tag{4}$$

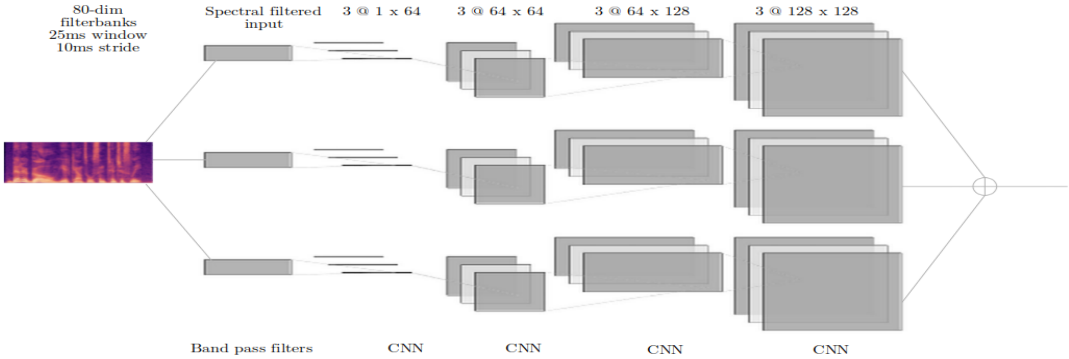


Fig. 1: Structure of multi-band CNN architecture: frequency filters applied in parallel through CNN layers

Put simply, the correlations between the neural representations will be higher when they have more similar information encoded within them.

To analyse the neural representations of the CNN layers and transformer layers, the activation embeddings of each neuron, at each epoch were extracted using a separately developed pipeline. To ensure consistency, this was done by passing a controlled input of 100 speech frames through each trained model, and extracting the activation output at each neuron. To aggregate the correlation coefficiency across layers, the spatial dimensions of the activation output vectors were flattened, which provided spatial representation of each neuron.

4. EXPERIMENTS

4.1. Data

The models were trained with the Switchboard dataset [18] with 300 hours of transcribed speech and evaluated on the Hub5'00 and RT03 test sets. 80-dimension filterbanks were extracted from 25ms windows with a stride of 10ms.

4.2. Multi-band CNN-Transformer

All models were compiled with the ESPRESSO framework [19]. The baseline model has a multi-layer stacked 2-dimensional CNN with pyramidal structure from 1 to 128 dimensions, with kernel size 3 x 3 and stride 1 and batch normalisation between each CNN layer [20]. The final convolutional layer is then projected to a transformer encoder-decoder model, described in [17]. The transformer model has stacked encoder layers with embedding dimensions of 512 x 2048 and 6 decoder layers with positional embeddings.

As can be observed in Table 1, the multi-band and multi-channel CNN models perform comparatively well to the baseline model on the Hub5'00 test sets despite not being fully optimised parameter-wise. The multi-band model achieves a lower word error rate (WER) on the Switchboard test set but slightly worse on all other test sets, while the multi-channel model performs slightly worse on both test sets. A multi-band model with dropout regularisation of 0.1 for each band was also included, in an attempt to improve the generalisation of the network. While this proved to improve the performance on the Callhome test set, the Switchboard set showed no improvement.

Table 1: CNN-transformer architectures performance on Hub5'00 Switchboard and Callhome test sets

Model	Swbd	Clhm
CNN + transformer	10.7	20.2
Mchan CNN + transformer	10.4	20.4
Mband CNN + transformer	10.5	20.5
Mband CNN + dropout + transformer	10.6	20.2

Table 2: CNN-transformer architectures performance on RT03 Switchboard and Fisher test sets

Model	RT03 S	RT03 F
CNN + transformer	21.2	13.3
Mchan CNN + transformer	23.5	15.2
Mband CNN + transformer	23.3	14.9
Mband CNN + dropout + transformer	23.5	15.5

Furthermore, as can be observed in Table 2, the performance of the multi-band and multi-channel approaches are both also worse on the RT03 test sets.

Figure 2 displays the WER over the validation set across epochs. It can be observed that, initially there is a large spike in WER on all models, although this is significantly higher on the baseline CNN-transformer model. The multi-band with dropout and multi-channel models had the smallest spikes in WER during training, which can be partly attributed to the regularisation effect of the dropout parameter. Despite the multi-band models displaying significantly more stability of error rates during training over the initial epochs, all models converged to roughly the same error rate by epoch 12.

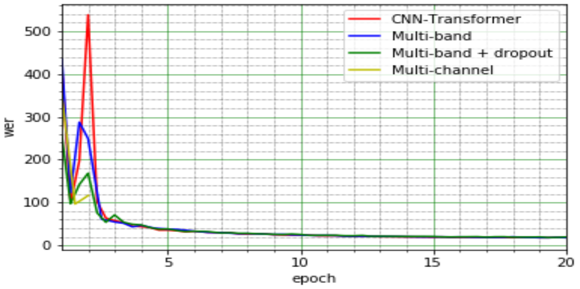
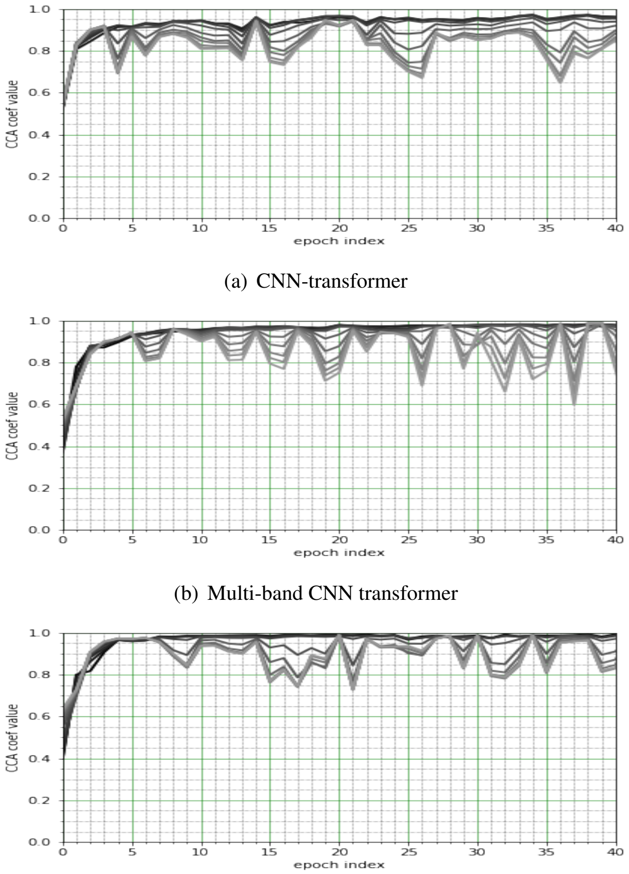


Fig. 2: Validation set word error rate across models during training on Switchboard data



(a) CNN-transformer

(b) Multi-band CNN transformer

(c) Multi-channel CNN Transformer

Fig. 3: Transformer models correlation coefficients through time as performance converges; the darker colour gradients are higher layer representations, while the lighter the gradient the deeper the layer.

5. DISCUSSION

The graph in Figure 3(a) shows the neural representation coefficiency across the layers of the baseline CNN-transformer model through training. There is a distinct hierarchical behaviour observed without clear convergence even in the earlier layers of the network, represented by the darker lines. The uncertainty of the attention mechanism has been highlighted by the pathology across the epochs, which could suggest that parameter re-weighting for the context information is occurring in the deeper layers, represented by the lighter lines. Very similar patterns can be observed through Figures 3(b) and 3(c) as both present the same instability of the deeper layers throughout the epochs during training. However, one of the only distinctions is that the convergence of the earlier layers appears to occur earlier, at epoch 20, with the multi-band model. Despite these small representational differences in the neural representations across the earlier layers of each model, they performed similarly on the Hub5'00 test set.

Furthermore, the experiment scenarios in this paper are set with CNNs, multi-channel CNNs, and multi-band CNNs to explicitly distinguish the input layers for the following encoder-decoder transformer layers. The multi-band CNN model explicitly modeled different frequency bands of the acoustic signal separately and aggregated them together. The multi-channel CNN model learned different representations of the same acoustic signal with a mixture of experts approach to aggregate these representations. Although there was a difference in convergence speed during training, the overall representation learning and performance remain similar on the Switchboard train set. Thus it can be hypothesised the transformer layers adapted different types of input representations to a similar average representational space, which highlights the memorisation capability of the transformers rather than the generalisation. The performance of the CNN, multi-channel CNN and multi-band CNN varied in the Callhome and RT03 test sets. As the models were trained with Switchboard data, if the transformer layers had generalised the input acoustic signal and the target categorical lexicon distribution mapping, the result patterns on other test sets should have been similar. These empirical results indicate that the transformers do more memorisation than generalisation with the training data.

6. CONCLUSION

Multi-band and multi-channel models have been implemented for an End-to-End ASR task, with comparable results to the baseline CNN-transformer approach on in-domain data but worse on the out-of-domain data. The analysis of neural representations within the models provided insight into the potential memorisation behaviour of the transformer architecture. An extension to this work would be the analysis of neural representations within End-to-End models on augmented or noisy data to observe the properties of different layers. These insights can be beneficial for few-shot learning model development in the ASR domain.

7. REFERENCES

[1] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Rwth asr systems for librispeech: Hybrid vs attention–w/o data augmentation," *arXiv preprint arXiv:1905.03072*, 2019.

[2] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[3] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin, "You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2020, pp. 439–444.

[4] M. Wiesner, A. Renduchintala, S. Watanabe, C. Liu, N. Dehak, and S. Khudanpur, "Pretraining by backtranslation for end-to-end asr in low-resource settings," *arXiv preprint arXiv:1812.03919*, 2018.

[5] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[6] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," *arXiv preprint arXiv:1402.1869*, 2014.

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[8] S. Wang, B. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv:2006.04768*, 2020.

[9] R. Masumura, N. Makishima, M. Ihori, A. Takashima, T. Tanaka, and S. Orihashi, "Hierarchical transformer-based large-context end-to-end asr with large-context knowledge distillation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5879–5883.

[10] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv preprint arXiv:2101.03961*, 2021.

[11] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[12] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," *arXiv preprint arXiv:2106.05974*, 2021.

[13] A. Ollerenshaw, M. A. Jalal, and T. Hain, "Insights on neural representations for end-to-end speech recognition," *Proc. Interspeech 2021*, pp. 4079–4083, 2021.

[14] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," *arXiv preprint arXiv:1706.05806*, 2017.

[15] E. M. Grais, F. Zhao, and M. D. Plumbley, "Multi-band multi-resolution fully convolutional neural networks for singing voice separation," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 261–265.

[16] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[19] Y. Wang, T. Chen, H. Xu, S. Ding, H. Lv, Y. Shao, N. Peng, L. Xie, S. Watanabe, and S. Khudanpur, "Espresso: A fast end-to-end neural speech recognition toolkit," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 136–143.

[20] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4845–4849.