This is a repository copy of *Probing statistical representations for End-to-End ASR*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/203758/

Version: Accepted Version

Proceedings Paper:
Ollerenshaw, A., Jalal, M.A. and Hain, T. orcid.org/0000-0003-0939-3464 (2023) Probing statistical representations for End-to-End ASR. In: 2023 31st European Signal Processing Conference (EUSIPCO) Proceedings. 2023 31st European Signal Processing Conference (EUSIPCO), 04-08 Sep 2023, Helsinki, Finland. Institute of Electrical and Electronics Engineers (IEEE) , pp. 401-405. ISBN 9789464593600

https://doi.org/10.23919/EUSIPCO58844.2023.10290070

© 2023 The Authors. Except as otherwise noted, this author-accepted version of a paper published in 2023 31st European Signal Processing Conference (EUSIPCO) Proceedings is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

PROBING STATISTICAL REPRESENTATIONS FOR END-TO-END ASR

Anna Ollerenshaw, Md Asif Jalal, Thomas Hain

Speech and Hearing Group, The University of Sheffield, Sheffield, UK

ABSTRACT

End-to-End automatic speech recognition (ASR) models aim to learn a generalised speech representation to perform recognition. In this domain there is little research to analyse internal representation dependencies and their relationship to modelling approaches. This paper investigates cross-domain language model dependencies within transformer architectures using SVCCA and uses these insights to identify critical parameters and improve recognition performance. It was found that specific neural representations within the transformer layers exhibit correlated behaviour which is related to recognition performance.

Altogether, this work provides analysis of the modelling approaches affecting contextual dependencies and ASR performance, and can be used to create or adapt better performing End-to-End ASR models without the requirement for hyperparameter optimisation, and also for downstream tasks.

Index Terms— speech recognition, end-to-end, cross domain, transformer, analysis, language modelling

1. INTRODUCTION

The typical approach to develop a framework for ASR has been to use deep neural networks to recognise and align acoustic features to graphemes or phonemes; replacing the requirement for distinctly-optimised modules, such as acoustic, pronunciation or language models (LMs). Using End-to-End modelling approaches reduces the need for expert domain knowledge as it aims to jointly optimise the training regime while adapting to diverse speech environments. These factors have led to End-to-End ASR models becoming a popular choice for on-device deployment.

Current research in End-to-End ASR modelling is dominated by three approaches: recurrent-transducers [1], Connectionist Temporal Classification models [2] or attention-based encoder-decoder architectures [3, 4]. End-to-End ASR frameworks are typically dependant on the amount of data resources and are commonly fine-tuned to the corpora in order to improve the recognition performance, which has directed techniques to improve the representation capacity of the modelling approaches [5, 6]. The ability to use larger amounts of training data significantly improves the recognition performance of End-to-End models [7]. However, previous work [8] suggested that the relationship between memorisation and generalisation within these networks remains elusive, being referred to as a "*black-box*". It can be hypothesised that richer neural representations are not analogous to increasing neural depth or model size [9]. By learning representations of speech that are robust, the general recognition performance of the model should improve proportionately without the requirement for increasing model size or training data.

This work was partly supported by Liveperson at the Liveperson Research Center

As the development of End-to-End ASR models continues, there is also a corresponding demand to be able to not only optimise the training process, but also to interpret the internal latent representations and the explainability of modelling approaches, such as the potential relationship between training data and the integration of LMs. Despite numerous variations of modelling approaches, there has also been little exploration of the internal model representations and their relationship to model recognition performance across different domains. Layer-wise analysis of models has been used to interpret modelling approaches and relationships between representations in multiple domains [10, 11, 12]. SVCCA analysis techniques have been used to highlight neural representations with respect to their ability to generalise, by observing the relationship between the correlation coefficients of neural layers during training [13].

As transformer modelling approaches achieve state-of-the-art results for End-to-End ASR, this work aims to identify and show representations in transformer models that adapt to out-of-domain LMs by analysing how the representations evolve across layers. The relationship between ASR performance and the neural representations is shown to aid parameter optimisation. Using unmatched sub-word LMs, it is possible to observe the dependencies of the representations within model layers and identify settings that improve model performance. Section 5 shows that observing the representation dependencies is important to develop intuitive modelling approaches and improve recognition performance.

The analysis methodology is first defined in Section 2, then the developed framework for analysis experiments is described in Section 3.3. The implementation of cross-domain LMs in the End-to-End ASR models is defined in Section 3.2 and the analysis experiments are shown in Section 4. Experiments analysing the adaptation of transformer model parameters are conducted in Section 5 with the results discussed in further detail in Section 6. It was found that deeper transformer model layers contain learned representation dependencies for cross-domain LMs and recognition performance can be improved by *tuning* the parameters to the hierarchical dependencies.

2. CORRELATION ANALYSIS METHODOLOGY FOR END-TO-END ASR MODELS

End-to-End ASR modelling approaches aim to approximate the mapping of an acoustic input $\mathcal{X} = \{x_1, ..., x_T\}$ of length $\mathcal{T}$ over output labels $\mathcal{Y} = \{y_1, ..., y_N\}$. The dependencies of End-to-End model parameters are not as tractable as approaches where acoustic and language information is modelled separately. These End-to-End or *"black-box"* models have inherently complex internal functions and it can be difficult to understand the relationship between the internal parameter dependencies and the resulting performance across datasets. Previous work in [14] showed it is possible to analyse the relationship between two sets of variables, such as neurons and their activation vectors by computing the correlations among the

eignenvectors of the covariance matrices of each set. This attempts to capture the direction of greatest variance and identify important neurons or activation vectors within the network. This method has been used to measure the linear relationship between vectors representing neural layers within ASR models [12, 13, 15]. The activation outputs of a layer $z_i^l = \{z_i^l(x_1), ..., (z_i^l(x_N)\}$ are extracted for dataset $\mathcal{X}$ and neuron i in layer l. To observe the similarity between or within neural layers and to compare the relationship of the internal representation dependencies upon different variables, the use of different LMs and model parameters are varied and the similarities between layer representations are compared to derive insights regarding their relationship.

To measure correlation for $\mathcal{N}$ datapoints, pairs of vectors are sampled from layers l_1 and l_2, which are projected, where $l_1 = \{z_1^{l_1}, ..., z_{N_1}^{l_1}\}$ and $l_2 = \{z_2^{l_2}, ..., z_{N_2}^{l_2}\}$. The projections of l_1 and l_2 are pruned by the application of singular vector decomposition (SVD) to retain 99% of the representative dimensions and to reduce the impact of potential noise. The application of SVD forms subspaces $l_1' \subset l_1$ and $l_2' \subset l_2$ and then CCA can then be applied to find vectors v and s that maximise correlation ρ between the projections l_1' and l_2':

$$\rho = \frac{\langle v^T l_1', s^T l_2' \rangle}{||v^T l_1'|| \, ||s^T l_2'||} \tag{1}$$

where v, s are transforms that aim to maximise the correlation of the vectors. ρ increases where neural representations have encoded more similar information. Further details regarding SVCCA can be found in [14].

3. EXPERIMENTAL SETUP

3.1. End-to-End ASR Model

Transformers, initially published in [3], are a widely chosen encoder-decoder architecture for speech recognition frameworks due their ability to parallelise the training regime. This enables use of larger amounts of training data which has been shown to improve recognition performance [7].

A CNN front-end is incorporated for feature extraction. The final convolutional layer is then projected to 12 stacked transformer encoder blocks with embedding dimensions of 512×2048 and 6 decoder layers with positional embeddings, compiled in the ESPRESSO framework [16].

3.2. Language Modelling in End-to-End ASR

The integration of LMs within End-to-End models can be used to supervise training optimisation and also for decoding to improve recognition performance [17, 16]. However, it is unclear how the internal dependencies of End-to-End models handle latent LM representations and whether there are similar learned representation spaces that are robust across different domains. By training models with cross-domain LMs, it is hypothesised to be possible to observe these dependencies using SVCCA analysis.

In the following experiments, a sub-word LM is integrated by shallow-fusion decoding [18] and label smoothing [19] techniques. The sub-word LM is a 3 layer LSTM model. Shallow-fusion decoding computes the weighted sum of a pair of posterior distributions over sub-words; using one from the ASR model and one from the sub-word LM. The sub-word LM is an LSTM-based LM trained with restricted computational complexity, by only keeping the most frequent sub-words and splitting the rest into characters, to enable conversion with low information loss. Label smoothing computes the cross entropy loss during the model's training regime with a weighted mix of distributions from a unigram LM and one-hot targets from the dataset.

3.3. Correlation Analysis Framework

The framework developed in [12] was utilised to investigate the relationships between internal dependencies. For all the experiments, the models were trained using the ESPRESSO framework [16]. The analysis was conducted for all models by extracting the activation outputs of each neural layer of the encoder for each training epoch. Each model was saved throughout all epochs and then a controlled input of 100 frames of unseen speech data was fed through the layers, whilst simultaneously extracting the activation outputs for each layer. 80-dimensional log Mel acoustic features with additional pitch features were extracted, from 25ms windows with a stride of 10ms.

3.4. Data

For the experiments, three common US-English datasets from differing domains for ASR were chosen: Switchboard [20], Librispeech [21] and WSJ [22]. The Switchboard dataset contains conversational telephone speech, Librispeech is a compilation of read audiobooks, and WSJ contains read news. The test sets for the Switchboard dataset, referred to as *Swbd* and *Callhome*, are derived from the LDC2002S09 set and contain 20 unreleased telephone conversations from Switchboard and 20 telephone unscripted conversations from Callhome. To ensure the transformer models were converged by training on comparative data to the LM, the model trained with Switchboard used up-sampled data (to 16kHz). The LMs trained with Librispeech were trained using the full 960 hour training set and the ASR models were tested on the *test-clean* and *test-other* sets. The training set for the WSJ LM was the *si284* set, with the *Dev93* set for validation and *Eval92* for testing the ASR model performance.

4. CROSS-DOMAIN LANGUAGE MODEL EXPERIMENTATION

Correlation analysis of the neural representations across the transformer model layers is used to measure and analyse the changes in correlation when cross-domain LMs are integrated. Figure 1 shows the SVCCA coefficients, as training converges, between the encoder layers of two transformer models. The models were trained with Switchboard data but one model uses an in-domain Fisher sub-word LM, and the other model uses an out-of-domain WSJ sub-word LM. These models are both trained with sub-word units using SentencePiece [23] and integrated during the training process using the scheduled sampling method and decoded with shallow-fusion, as described in Section 3.2. The correlation analysis shows very little difference in coefficiency between layers 1 to 6 (top graph of Figure 1), aside from in the initial epochs which could be attributed to the random initialisation of parameters. This suggests that the neural layers of both of these models are converging to similar representation spaces. However, between layers 7 to 12 (bottom graph of Figure 1, the differences in coefficiency are much larger throughout training. This suggests that the representations learned in these deeper layers are more dependent upon the LM domain.

The bottom graph in Figure 2 displays the standard deviations of the coefficiency between the models trained with cross-domain LMs. This aims to show the variation in coefficiency by layer more clearly, where the standard deviations in layers 10, 11 and 12 are

highest. The top graph of Figure 2 shows the variance in coefficiency within the neural layers of a model trained without scheduled sampling or shallow-fusion decoding compared to the model trained with the Fisher sub-word LM. This suggests that a similar observation can be made for LM specific representations, whereby the variance is higher overall and the coefficiency of layers 8 to 12 deviates the most. The results in Figure 2 also imply that layers 1 to 4 have very little dependency on LM representations. These insights suggest that encoder layers 1 to 4 of the transformer model can be frozen or stopped early when fine-tuning with LMs and the optimisation regime of End-to-End ASR models can be adapted to improve downstream tasks.

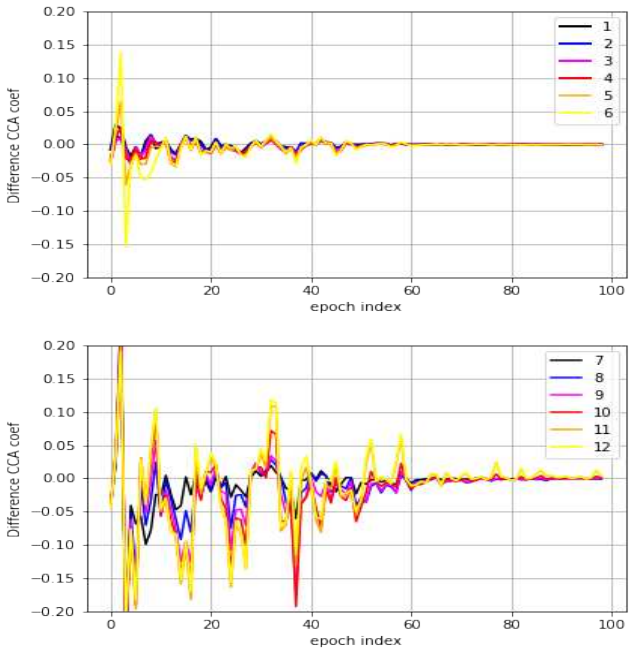

Fig. 1: SVCCA correlation coefficients as performance converges within transformer layers 1 to 6 (**top**) and layers 6 to 12 (**bottom**), between a model trained with a Fisher-based LM and a model trained with a WSJ-based LM

Regarding performance, the model that was trained with the Fisher LM reached 9.5% word error rate (WER) on the Switchboard test set and 19.1% on the Callhome test set, while the model that was trained with the WSJ LM was 10.7% and 21.1% respectively. The differences in recognition performance are attributed to the domains of the LMs and the test sets used for evaluation.

5. MODELLING STRUCTURE ANALYSIS

To optimise the parameters for state-of-the-art End-to-End ASR models, many iterations are trained with parameter modifications. Optimisation of model parameters to specific datasets to achieve the best recognition performance possible [16, 17] is referred to here as *tuned*. For example, the dimensionality, number of layers and also the hyperparameters have been observed to impact the recognition performance. As shown in Table 1, using a transformer model with the same parameters and composition for several datasets does not achieve the lowest WER across all of the datasets. These *tuned* models are reached by extensive hyperparameter optimisation techniques, which are computationally expensive and considerably time consuming without providing observational evidence regarding the dependencies of certain parameters upon the recognition performance.

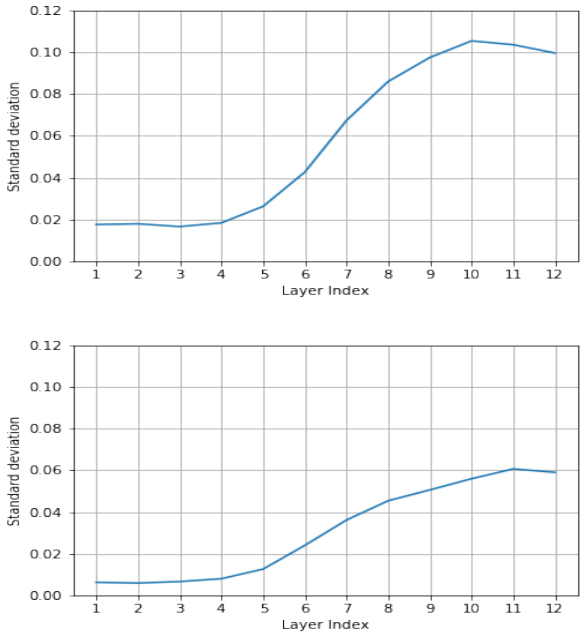

Fig. 2: Standard deviation of correlations across transformer model layers with and without a LM (**top**) and with cross-domain LMs (**bottom**)

Using cross-corpora correlation analysis, it is possible to interpret the dependencies of parameters in a more meaningful way and provide some observational evidence to reduce the need for hyperparameter optimisation when developing new models or fine-tuning trained models. By understanding the representation dependencies, it is possible to identify which parameters are unlikely to improve model performance, which can potentially reduce the computational resources required. Table 1 shows the results of 3 transformer models with variations in model parameters that are used in state-of-the-art End-to-End ASR frameworks. All models are the same transformer-based encoder-decoder architecture with the following variations:

- **Model 1** has an embedding dimension of 512, a feed forward embedding dimension of 2048, 4 attention heads, and an attention dropout of 0.25.
- **Model 2** has an embedding dimension of 256, a feed forward embedding dimension of 1024, 4 attention heads, and attention dropout of 0.25.
- **Model 3** has an embedding dimension of 512, a feed forward embedding dimension of 2048, 8 attention heads, and an attention dropout of 0.1.

To observe the relationship between the learned representations of the adapted models and attribute these adaptations to improved recognition performance with specific data, the model performance was assessed across all test sets, as shown in Table 1. For the Switchboard and Callhome test sets, the recognition performance of model

1 is the best, while model 2 reaches slightly worse performance on the Callhome set and model 3 has the highest WER for both test sets.

Table 1: Transformer model WER on EVAL'00, WSJ and Librispeech test sets with *tuned* parameters

Model	Swbd	Chm	Eval92	Dev93	Test-cln	Test-oth
M1	**9.5**	**19.1**	4.59	7.54	3.5	8.51
M2	9.6	20	**4.13**	**6.3**	3.99	8.72
M3	10.4	21.6	4.52	7.43	**1.9**	**3.9**

Figure 3 displays the SVCCA coefficients for each model trained with the Switchboard dataset. Model 2's mean coefficiency, across layerwise representations, are substantially less correlated than the other models. The standard deviations of the correlations within these layers also vary significantly higher than Model 1 or 3. Model 3's mean coefficiency across layerwise representations is fairly similar for all layers with very small standard deviation. It is observed that correlations within the layers of model 2 have lower coefficiency, and the recognition performance of this model is lower for the Switchboard test sets. Also, there are little hierarchical coefficiency patterns throughout the layers of model 3, and this model also has a slightly worse performance, which corroborates with results from [12]. Model 1 has lower coefficiency within layers 8-12 and has the best recognition performance.

6. DISCUSSION

The findings in Section 3.2 correlate with findings from [24] where semantic and syntax level features of speech are predominantly dependent upon deeper layers of transformer-based models, while acoustic and fluency features are predominantly dependent on the shallower layers. In the case of End-to-End transformers for ASR tasks, the LM-dependent representations are shown to be primarily dependent within layers 7-12. The cross-domain LM-dependent representations are observed within layers 10-12. Further experiments training with the WSJ dataset with cross-domain sub-word LMs showed very similar observed behaviours across layer coefficiency. These observations can be used to identify the elements that affect recognition performance, without the need for extensive training requirements, and to improve joint optimisation. The analysis also aids in the identification and interpretability regarding the representation dependencies within End-to-End ASR models.

The experiments in Section 5, attempt to show these internal dependencies with regard to the model parameters within the same modelling architectures. As shown in Figure 3, Model 2 used shallower embedding dimensions than model 1, which has caused the coefficiency of many of the layers to become highly uncorrelated. Model 3 is observed to have very highly correlated layers, however there are little distinct hierarchies in the neural representations when the attention heads are increased to 8 and the attention dropout is reduced. By adapting the parameters of transformer models, the layers with the most dependency for representing domain-specific information are altered. These changes in hierarchical representations have been observed to impact recognition performance, and further suggests a relationship between correlated hierarchical representations and the ability for the model to generalise, particularly for cross-domain speech recognition. Increasing the attention dropout is theorised to improve model robustness [6], where typical features of conversational speech are boundary uncertainties and hesitations. In the case of End-to-End conversational speech recognition, the results

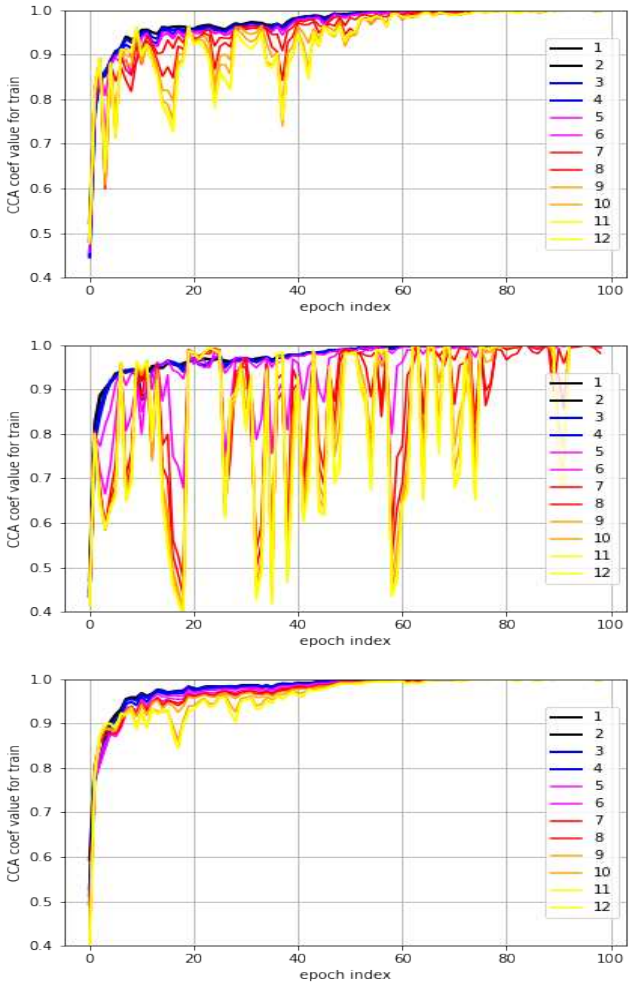

Fig. 3: Transformer SVCCA coefficients as performance converges in model 1 (top), model 2 (middle) and model 3 (bottom) trained with Switchboard data

show that using substantial attention dropout in transformer models is important to produce correlated hierarchies in dependent layers but also utilise a model with sufficient embedding dimensionality that the representations within context-critical layers don't become too uncorrelated.

7. CONCLUSION

Using SVCCA as a correlation index has identified several aspects of the relationships between the neural representations, transformer-based modelling parameters and the impact these have upon recognition performance. Interpretative analysis is important to develop future modelling approaches for meaningful improvement strategies. Expanding the scope of the investigation into the attributes and potential learned features that could be classified within the layers would provide a deeper understanding of the properties of these dependencies and how these could be further exploited. The insights into the dependencies of the neural layers can be used for the development of models for few-shot learning and downstream tasks for End-to-End ASR.

8. REFERENCES

[1] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[2] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[5] Chanwoo Kim, Kean K Chin, Michiel Bacchiani, and Richard M Stern, "Robust speech recognition using temporal masking and thresholding algorithm," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[7] Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Rwth asr systems for librispeech: Hybrid vs attention–w/o data augmentation," *arXiv preprint arXiv:1905.03072*, 2019.

[8] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al., "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.

[9] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio, "On the number of linear regions of deep neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[10] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton, "Similarity of neural network representations revisited," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3519–3529.

[11] Ari Morcos, Maithra Raghu, and Samy Bengio, "Insights on representational similarity in neural networks with canonical correlation," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[12] Anna Ollerenshaw, Md Asif Jalal, and Thomas Hain, "Insights on neural representations for end-to-end speech recognition," *Proc. Interspeech 2021*, pp. 4079–4083, 2021.

[13] A Ollerenshaw, MA Jalal, and T Hain, "Insights of neural representations in multi-banded and multi-channel convolutional transformers for end-to-end asr," in *IEEE 30th European Signal Processing Conference*. Institute of Electrical and Electronics Engineers (IEEE), 2022.

[14] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," *Advances in neural information processing systems*, vol. 30, 2017.

[15] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[16] Yiming Wang, Tongfei Chen, Hainan Xu, Shuoyang Ding, Hang Lv, Yiwen Shao, Nanyun Peng, Lei Xie, Shinji Watanabe, and Sanjeev 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.

[17] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[18] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

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

[21] 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.

[22] Douglas B Paul and Janet Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[23] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[24] Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.