What Happens To WavLM Embeddings During Speech Emotion Recognition Fine-tuning?

Anonymous submission to Interspeech 2024

Abstract

We investigate the dynamics of WavLM embeddings during fine-tuning for speech emotion recognition (SER). Leveraging the CREMA-D dataset, we fine-tuned a pre-trained WavLM model and analyzed the transformations in its embeddings. This involved examining both the acoustic features (specifically, eGeMAPS features) and the embeddings extracted at different stages of fine-tuning and from various transformer layers within the model. Our methodology included (i) training linear classifiers on the embeddings to evaluate their SER performance and (ii) utilizing Centered Kernel Alignment (CKA) to compare embeddings across fine-tuning phases and layers. Our findings suggest that fine-tuning affects the encoding of acoustic features to differing degrees, but with no catastrophic forgetting of acoustic phenomena. Additionally, our analysis elucidates how these changes are distributed across the model and the fine-tuning process, as well as their impact on the SER performance.
Index Terms: speech emotion recognition, self-supervised learning, acoustic features, interpretability

1. Introduction

Self-supervised speech models rely on pretext tasks to learn from unlabeled large-scale datasets and produce robust general-purpose speech representations [1]. These representations enable impressive results across various audio tasks, including speech emotion recognition (SER) [2, 3]. SER consists of detecting emotional states, focusing on the way something is said, rather than what is being said [4]. This involves categorizing emotions into specific groups, often using the *Big Six* framework—anger, fear, disgust, joy, sadness, and surprise [5].

Prior literature review on self-supervised SER models [6] shows that (i) fine-tuning pre-trained models considerably improves performance compared to using them as simple speech representation extractors for feeding linear classifiers; (ii) fine-tuning the transformer layers of a model and employing average time-pooling with a linear classifier on top typically yields state-of-the-art performance. Still, the explainability of what kind of information is contained in the self-supervised encoding spaces remains a challenge [1]. A recent study [7] sought to understand how acoustic information (i.e., entropy, jitter, shimmer, zcr, voicing, pitch, centroid, duration) is encoded by these models. They fine-tuned pre-trained models for detecting emotional properties in a dimensional framework (considering arousal, valence, and dominance) and probed them to predict acoustic features. The hypothesis was that if fine-tuning enhances the model's ability to predict certain acoustic features, it suggests the model learns relevant information for the SER task from these features. However, the study found no significant changes in the encoding of information, except for a reduced emphasis on audio duration in the fine-tuned model.

While recent efforts have examined how acoustic information is encoded in speech representations generated by pre-trained self-supervised models, comparatively little is understood about how these representations change when models are adapted to solve the speech emotion recognition downstream task. Drawing inspiration from prior research [8] in the natural language processing (NLP) field, which checked the changes to BERT embeddings during task-specific fine-tuning. Our study aims to uncover the effects of SER fine-tuning on a self-supervised model representation. We investigate:

RQ1. How does self-supervised model fine-tuning for SER affect the encoding of acoustic features?

RQ2. How are changes in the self-supervised model distributed during this fine-tuning process?

RQ3. What are the implications of these modifications for the performance of speech emotion recognition?

We develop a pipeline[1] to analyze the encoding of interpretable, hand-crafted features within self-supervised models, particularly focusing on their significance for specific downstream tasks. Secondly, we provide an examination of how fine-tuning the self-supervised WavLM model [9] for SER impacts the encoding of the extended Geneva Minimal Acoustic Parameter Set (eGeMAPS) features [10].

2. Methodology

2.1. Speech representations

2.1.1. Data-driven embeddings

We used WavLM [9] as a self-supervised model for generating data-driven speech representations. Specifically, we employed the "large" checkpoint[2] which includes 317 million parameters and was pre-trained to English by the original model creators. We chose it for its leading performance in the SUPERB benchmark[3] at the time of experimentation, particularly in SER.

2.1.2. Acoustic features

We employed the eGeMAPS features [10] as our hand-crafted speech representation, which includes 88 parameters covering a range of quality-, vocal tract-, and spectrum-related properties, as it was designed to provide a comprehensive yet minimalistic representation of relevant paralinguistic information. The extraction of eGeMAPS features was facilitated by the OpenSMILE toolkit [11].

[1] https://github.com/github_to_come

[2] https://huggingface.co/microsoft/wavlm-large

[3] https://superbbenchmark.org/leaderboard

2.2. Dataset

We selected *CREMA-D* [12] due to its language compatibility with WavLM, its high degree of alignment with the canonical *Big Six* emotions, and its balanced distribution of emotions and sentences. Its diverse pool of 91 actors (43 females, 48 males, aged 20-74, from various self-identified racial and ethnic backgrounds) and 7,442 audio-visual samples offer a comprehensive yet manageable resource, superior to smaller datasets like SAVEE [13] and more computationally feasible than larger ones like MSP [14]. Additionally, the dataset's reliability is ensured through rigorous validation by 2,443 raters, ensuring accuracy and fair emotion representation.

2.3. Experimental setup

2.3.1. Data preparation

First, we used *pyloudnorm* [15] to normalize the CREMA-D loudness to -23 dB LUFS to mitigate its impact on the speech emotion recognition task. Although loudness may play a role in discerning emotions, it is susceptible to confounding variables, such as the distance from the microphone [16].

Next, we partitioned the CREMA-D dataset through a two-phase 80-20 split, resulting in three distinct subsets, hereinafter referred to as subset 1, subset 2, and subset 3. We adopted a stratified sampling approach ensuring balanced gender representation within each subset. We also maintain complete speaker independence among subsets by excluding any actor included in one subset from the others. This process yielded subsets of approximately 64%, 20%, and 16% of the entire dataset, corresponding to 4,662, 1,552, and 1,231 samples, respectively.

2.3.2. Model fine-tuning procedure

We fine-tuned the WavLM checkpoint for SER using the CREMA-D subsets 1 and 2 for training and validation, respectively. We implemented global average pooling across the time dimension of the final hidden layer to handle audio clips of various lengths, followed by a single linear layer for predicting the emotion category. This approach is in line with the strategies suggested in [6].

The fine-tuning process was carried out on an A100 GPU, selecting a batch size of 16 due to limitations in computational resources and available memory. The training spanned 50 epochs with a default learning rate of 1e-5, employed cross-entropy loss, and incorporated early stopping based on validation loss to mitigate the risk of overfitting. Early stopping criteria were implemented to terminate fine-tuning if at least a 0.025 improvement in loss was not observed after 50 training steps (equivalent to 5 evaluations or approximately 1.4 epochs). This measure, substantiated by preliminary analysis and evidenced by the training progression plot, was aimed at enhancing computational efficiency and mitigating the risk of overfitting. Periodic state-saving was implemented, where a checkpoint was automatically saved after every 50 training steps to gain insights into the model's evolution subsequently.

2.3.3. Speech representation similarity

To address RQ1 and RQ2, we utilized a state-of-the-art technique designed for analyzing the similarity in data representations. Centered Kernel Alignment (CKA) [17] is a similarity metric based on dot products which has been proven robust for linear assessing similarities in neural network representations [17]. The range of the CKA scores is from 0, indicating no similarity, to 1, indicating identical representations.

For each transformer layer in every checkpoint saved during the SER fine-tuning, we extracted average time-pooled embeddings for all clips in Subset 3. We also extracted the eGeMAPS features for the same data. Next, we standardized these speech representations and computed the CKA scores among them.

2.3.4. Layer-wise speech emotion recognition

To respond to RQ3, we trained linear classifiers on the embeddings from each transformer layer and checkpoint to predict the emotion category. We chose a simple linear classifier because if it can accurately identify emotions, it means emotional information is clearly encoded in the speech representations. We trained using an 80/10/10 split of subset 3 and assess their performance using F1 score. To minimize the effects of the linear layer's initialization, we train 10 different seeds and average their final F1 scores. For a comparative analysis, we replicated the above training-testing process using acoustic features. Instead of embeddings, we used the extracted eGeMAPS features to train a linear classifier for SER as in the previous step.

The training process was carried out on an A100 GPU, with a batch size of 16. The training spanned 1000 epochs (PyTorch Lightning Trainer's default) with a learning rate of 1e-2, employed cross-entropy loss, and incorporated early stopping based on validation loss. Early stopping was set to halt training if no loss improvement of at least 0.025 was seen after 5 epochs.

3. Results

The WavLM model underwent fine-tuning on CREMA-D subset 1 for approximately 9.6 epochs. The fine-tuning process was halted due to early stopping criteria being met. During fine-tuning, 35 evaluations were performed, resulting in the preservation of 7 checkpoints in addition to the original checkpoint, totaling 8 checkpoints overall.

Figure 1 illustrates the CKA similarity of the speech representations obtained from subset 3 across all transformer layers from each of the eight checkpoints in comparison with eGeMAPS acoustic features.

Figure 2 presents the SER performance on the test set generated on the subset 3 split, measured in terms of F1 score.

4. Discussion

4.1. Research Questions Revisited

[RQ1.] How does self-supervised model fine-tuning for SER affect the encoding of acoustic features?

The comparison of eGeMAPS acoustic features with the speech representations generated by different transformer layers of the WavLM checkpoints obtained during finetuning, as quantified by the CKA similarity metric, offers insights into how embeddings can describe acoustic phenomena. The similarity scores, ranging from 0.015 to 0.040, show that WavLM embeddings primarily do not encode acoustic information. However, the focus of this study is on the changes in acoustic information encoding during the SER fine-tuning process and not on their overall magnitude. Based on the findings illustrated in Figure 1, it can be concluded that the original acoustic information is preserved within the model, but there have been modifications in how this information is encoded (see RQ2).

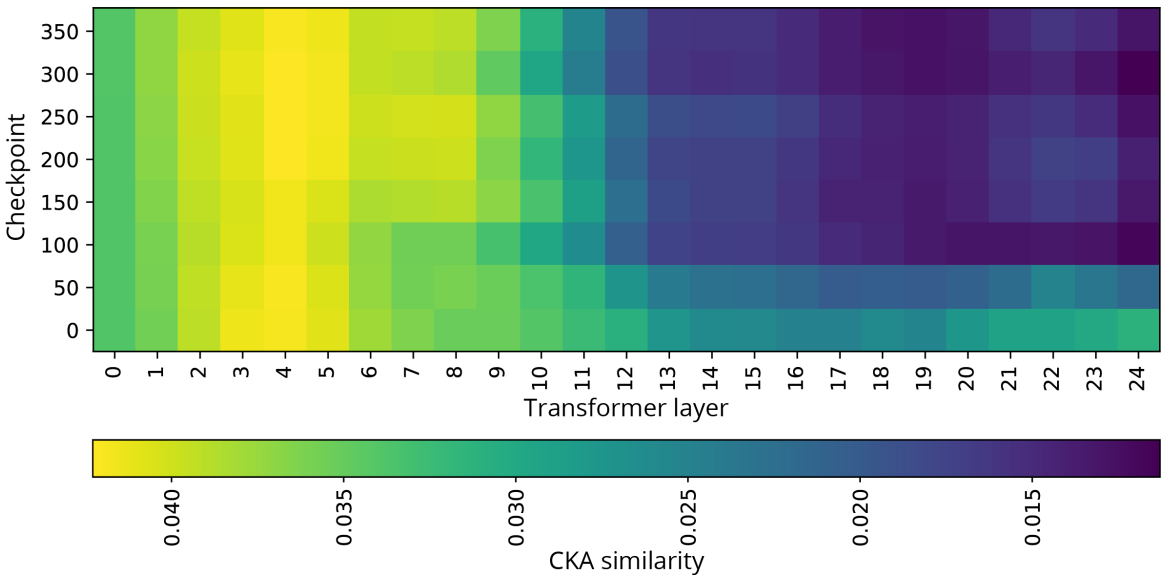

Figure 1: *The plot illustrates the Centered Kernel Alignment (CKA) similarity metric applied to the speech representations obtained from subset 3. The heatmap shows the comparison across all transformer layers at eight different checkpoints, with reference to the eGeMAPS acoustic features. Each cell represents the CKA similarity score between the transformer layer's outputs and the eGeMAPS features, where the horizontal axis corresponds to the 25 transformer layers and the vertical axis represents the checkpoints during fine-tuning, listed from 0 to 350 (number of steps during training). The color gradient ranges from yellow (higher similarity) to purple (lower similarity), indicating how the internal representations of the model align with the acoustic feature set over the course of fine-tuning.*

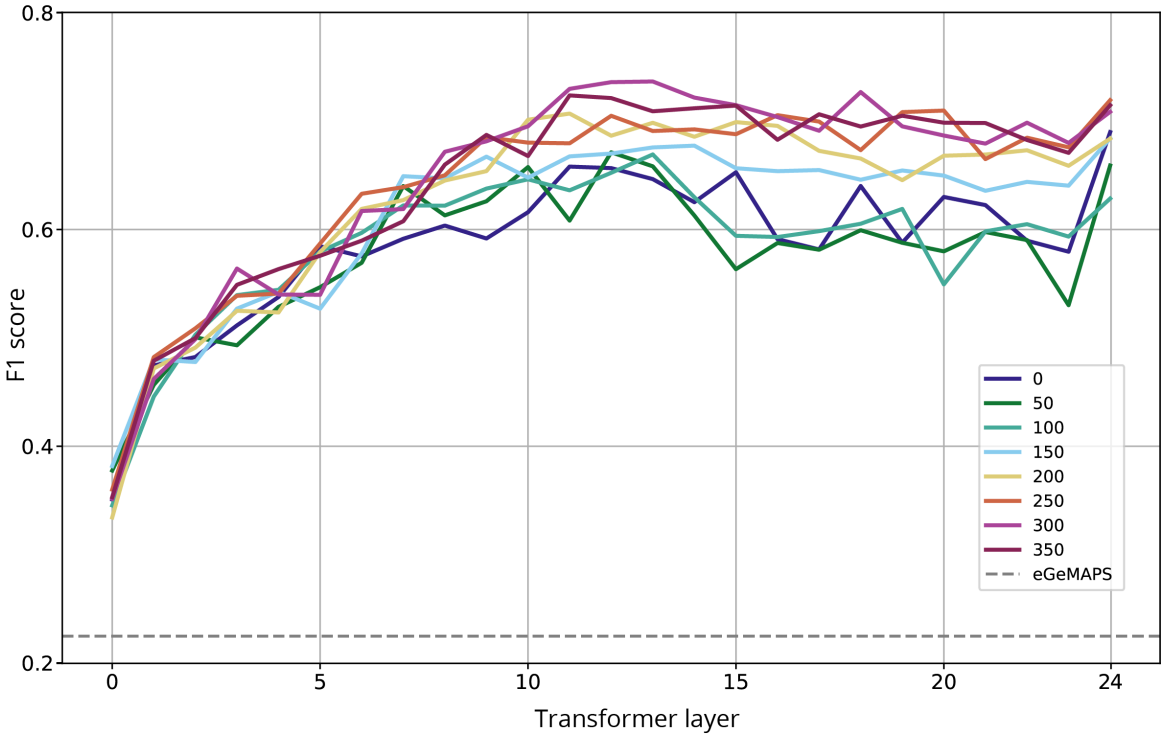

Figure 2: *The plot displays the F1 score-based SER performance on the CREMA-D dataset subset 3, comparing the outputs from different transformer layers at eight SER fine-tuning checkpoints (shown in various colors from blue for initial to red for the final checkpoint) against the baseline eGeMAPS features (dashed line). Each point is averaged over 10 seeds to lessen the impact of the linear classifier's initialization. This succinctly demonstrates the evolution and effectiveness of transformer-derived speech representations for emotion recognition.*

[RQ2.] How are changes in the self-supervised model distributed during this fine-tuning process?

Figure 1 reveals that acoustic information is primarily encoded in the early transformer layers of the original WavLM checkpoint (checkpoint 0).

The same Figure 1 shows how SER fine-tuning produces zone-specific changes rather than uniform adjustments. The trend of preserving acoustic information in early layers is maintained through fine-tuning, highlighting their role in capturing basic acoustic features. In contrast, the encoding of acoustic information in later layers diminishes as the model adapts to the SER task. This indicates a shift towards hierarchical information processing, where early layers focus on generic patterns, and later layers specialize in task-specific features. This pattern mirrors observations in deep NLP [8] and computer vision models [18], where early layers capture universal features (e.g., edges or linguistic patterns), and later layers focus on more complex, task-relevant patterns (e.g., facial features or specific linguistic contexts).

The hypothesis is that during fine-tuning, the network increases its reliance on features crucial for predictions, maintains information on adequately represented features, and reduces information on potentially detrimental features. This effect is particularly pronounced in the last layers since the fine-tuning loss is typically calculated based on the last layer's predictions, with weight modifications applied backward from there. The reduced emphasis on acoustic information in the later layers of WavLM during SER fine-tuning implies a prioritization of non-acoustic features for the SER task. Future research is needed to understand which information types are maintained and increasingly encoded in the later layers during SER fine-tuning. Previous research [6] indicates that explicitly incorporating linguistic information into self-supervised models for predicting emotional attributes does not enhance performance for arousal and dominance. It only sometimes benefits valence prediction. This lack of improvement suggests that the models might already possess the necessary linguistic information. This insight could support the hypothesis that linguistic information is the primary data utilized in SER tasks. Moreover, our finding prompts a reevaluation of the SER task, especially in light of previous attempts to identify acoustic correlates of emotion perception, which have produced inconclusive results.

[RQ3.] What are the implications of these modifications for the performance of speech emotion recognition?

Overall, classifiers using WavLM checkpoints perform better than those using eGeMAPS features (Figure 2). This suggests that WavLM generally captures more relevant information than simple acoustic features. Furthermore, early transformer layers show similar performance and are outperformed by later layers. This observation, coupled with the fact that later layers encode less acoustic information, particularly in fine-tuned checkpoints (see RQ1), challenges prior models of SER that relied heavily on acoustic correlates [10].

In examining the effects of fine-tuning on the performance of SER, we note an enhancement in performance at later checkpoints for the later transformer layers, obtaining state-of-the-art results[4], as we expected from literature guidelines [6]. Importantly, a plateau in the F1 score is observed around layer 12. This finding corroborates the insights in [6] in the context of similar speech self-supervised models (e.g., HuBERT). Specifically, this work suggests that reducing the number of transformer layers to 12 does not considerably compromise performance regarding the emotional characteristics of speech. It is noteworthy that the 12th layer marks the initial point where acoustic information encoding diminishes in the subsequent checkpoints (see Figure 1).

[4]https://www.paperswithcode.com/sota/speech-emotion-recognition-on-crema-d

4.2. Limitations

This study was conducted using the CREMA-D [12], which contains acted emotional speech, not naturalistic expressions. This raises questions about the generalizability of our results. Despite this limitation, our pipeline is designed to be adaptable to diverse datasets and models, potentially broadening the scope of future research and validation efforts.

In evaluating speech representation similarity, we employed the CKA metric with only linear predictors. However, the choice of a suitable similarity metric is debatable, with no agreed-upon best option. Other metrics like Canonical Correlation Analysis (CCA) [19] might show deeper insights as although CKA performs well in specificity tests, it lacks in sensitivity tests, where CCA excels [20] . This underscores the need for exploring diverse metrics to understand speech representation similarity better. Moreover, our method does not directly compare with [7] regarding the preservation of acoustic features like entropy, jitter, shimmer, ZCR, voicing, pitch, centroid, and duration during SER fine-tuning. This is because, unlike the probing approach used in [7], which predicts acoustic features from speech representations, our method does not check for the encoding of specific hand-crafted features in the self-supervised model. However, our approach should be viewed as complementary, rather than inferior, to that probing technique. Indeed, our method avoids potential biases that may arise from choosing and/or overfitting linear models for feature prediction.

In our layer-wise speech emotion recognition, we opted for a basic linear classifier, believing that accurate emotion identification with such a classifier would indicate a clear encoding of emotional information in speech representations. Recent literature, however, suggests that incorporating nonlinear layers, such as ReLU, reveals more nuanced details within these representations. Exploring more sophisticated models to delve deeper into this aspect presents a promising direction for future research.

5. Conclusion

In this study, we introduced a methodology to explore the impact of task-specific fine-tuning on interpretable, hand-crafted information encoded in self-supervised speech models.

Applying our method to the WavLM model in the context of SER, we found how fine-tuning progressively modifies a portion of the model's capacity, mainly affecting the later transformer layers by diminishing their encoded acoustic information. Our results demonstrate that classifiers leveraging embeddings with reduced acoustic information can achieve state-of-the-art performance in SER. These findings challenge existing SER models that primarily depend on acoustic correlates, suggesting a new direction for SER problem framing.

For future research, we propose extending our methodology to identify specific non-acoustic features the model utilizes for prediction, starting with linguistic features. Additionally, we aim to investigate which acoustic features are retained during fine-tuning using probing techniques.

6. References

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *arXiv preprint arXiv:2205.10643*, 2022.

[2] J. Turian *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[3] S. W. Yang, P. H. Chi, Y. S. Chuang, C. I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi *et al.*, "Superb: Speech processing universal performance benchmark," in *INTERSPEECH 2021*. ISCA, 2021, pp. 3161–3165.

[4] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[5] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, pp. 169–200, 1992.

[6] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[7] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, and B. W. Schuller, "Probing speech emotion recognition transformers for linguistic knowledge," in *Interspeech 2022*. ISCA, sep 2022. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2022-10371

[8] A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney, "What happens to bert embeddings during fine-tuning?" *arXiv preprint arXiv:2004.14448*, 2020.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[12] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[13] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition." in *AVSP*, vol. 2009, 2009, pp. 53–58.

[14] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Interspeech 2020*, 2020.

[15] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES Convention*, 2021.

[16] G. Zhang, S. Qiu, Y. Qin, and T. Lee, "Estimating mutual information in prosody representation for emotional prosody transfer in speech synthesis," in *ISCSLP 2021*. IEEE, 2021, pp. 1–5.

[17] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *International conference on machine learning*. PMLR, 2019, pp. 3519–3529.

[18] K. Ohri and M. Kumar, "Review on self-supervised image recognition using deep neural networks," *Knowledge-Based Systems*, vol. 224, p. 107090, 2021.

[19] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," 2017. [Online]. Available: https://arxiv.org/abs/1706.05806

[20] F. Ding, J.-S. Denain, and J. Steinhardt, "Grounding representation similarity with statistical testing," *arXiv preprint arXiv:2108.01661*, 2021.