Pitfalls in Measuring Neural Transferability

Suryaka Suresh SURYAKAS@IIITD.AC.IN

Vinayak Abrol ABROL@IIITD.AC.IN

CSE Department & Infosys Centre for AI, IIIT Delhi, India

Anshul Thakur ANSHUL.THAKUR@ENG.OX.AC.UK

Department of Engineering Sciences, University of Oxford, UK

Editors: Sophia Sanborn, Christian Shewmake, Simone Azeglio, Nina Miolane

Abstract

Transferability scores quantify the aptness of the pre-trained models for a downstream task and help in selecting an optimal pre-trained model for transfer learning. This work aims to draw attention to the significant shortcomings of state-of-the-art transferability scores. To this aim, we propose *neural collapse-based transferability scores* that analyse intra-class *variability collapse* and inter-class discriminative ability of the penultimate embedding space of a pre-trained model. The experimentation across the image and audio domains demonstrates that such a simple variability analysis of the feature space is sufficient to satisfy the current definition of transferability scores, and there is a requirement for a new generic definition of transferability. Further, building on these results, we highlight new research directions and postulate characteristics of an ideal transferability measure that will be helpful in streamlining future studies targeting this problem.

Keywords: Transfer learning, acoustic modelling, neural networks, model explainability

1. Introduction

Deep learning has shown remarkable performance across multiple domains, including image, audio and speech processing tasks Goodfellow et al. (2016); Amodei et al. (2016); Abeßer (2020). The success of deep learning can be attributed to the availability of a large amount of training data. However, many classification tasks (such as in bioacoustics and healthcare) often suffer from the scarcity of the labelled training data and hence, deep neural networks (DNNs) trained in data-scarce scenarios are usually prone to over-fitting Thakur et al. (2019); Viksit and Abrol (2023). Although recent advances in few-shot or deep metric learning Schroff et al. (2015) have been made to address this problem, the generalisation capability of such few-shot models has recently been shown to be questionable Wang et al. (2020).

In contrast, transfer learning allows a neural network to exploit the weak prior knowledge obtained from other tasks to learn new or downstream tasks Van Den Oord et al. (2014); Nguyen et al. (2020a). This weak prior knowledge is "embedded" in the parameters of the pre-trained models, and it might be possible to adapt or fine-tune these parameters to learn downstream tasks using only a handful of training examples Thakur et al. (2022). Moreover, few-shot learning frameworks can also benefit from exploiting weak prior knowledge in the form of pre-trained models Wang et al. (2020); Pons et al. (2019). The first challenge in exploiting transfer learning for a downstream task is to select an appropriate model from a large collection of candidate pre-trained models. For example, predicting COVID-19 from X-ray images of the lungs presents the challenge of selecting a CNN model that could be pre-trained on natural images or other X-ray images. However, the choice of an optimal pre-trained model and, hence, the appropriate pre-trained tasks or modalities has been a subject of contention in multiple domains such as audio analysis. Some studies have used CNNs trained on natural images to classify audio spectrograms Zhou et al. (2018); Shin et al. (2021). Whereas the other studies have focused on transferring knowledge from audio-based pre-trained models for capturing richer domain or task-specific richer characteristics Koike et al. (2020). This suggests that existing attempts for optimal pre-trained model selection for transfer learning are mainly anecdotal. As a result, there is a requirement for transferability scores that can quantify the appropriateness of a pre-trained model for a given task without training/fine-tuning the model itself.

Although existing studies have shown some success in the optimal ranking of pre-trained models for classification tasks, we argue that these methods essentially oversimplify the concept of *neural transferability* to indirectly measure the discriminative nature of embedding from a pre-trained model. This oversimplification is justified if we are using pre-trained models as feature extractors to train a classifier for the downstream task. However, in most transfer learning settings, such as re-training with pre-trained weights or freezing/re-training some pre-defined layers, this oversimplification is not enough to correctly identify the optimum pre-trained models. This is mainly because the model complexity and the sample complexity are not taken into consideration while defining such scores.

To further shed light on pitfalls of state-of-the-art neural transferability scores, this paper introduces *neural collapse* based transferability scores to obtain essential insights into the resonance of a pre-trained model with the downstream classification task. Neural collapse defines the behaviour of a DNN during the terminal stage of its supervised training with cross-entropy loss Papyan et al. (2020). It characterises the presence of a simple geometric structure, i.e. class-specific variability collapse among penultimate layer embedding of downstream training examples. We argue that this can also be used to infer a model's suitability for the intended downstream task. In other words, the proposed scores express the degree of neural transferability in terms of the discriminative nature of DNN's embedding space. Experimental results highlight that the proposed score, while being simple and efficient, performs either comparable or better than existing methods.

The success of the proposed neural collapse scores upheld the previously mentioned oversimplification claim in existing state-of-the-art and questions their appropriateness for wider deep transfer learning. Based on these observations, this paper also postulates characteristics of an ideal transferability measure, such as sensitivity towards model & data complexity and support for the wider downstream tasks (data generation, image/audio segmentation and clustering). The realisation of this postulated transferability score can have wider implications for different and less explored frontiers of applied transfer learning, such as healthcare informatics and accelerated drug discovery. The major highlights of this paper are as follows:

- Deviating from anecdotal reasoning, this paper introduces a neural collapse-inspired quantification measure to select the optimal pre-trained model for transfer learning.

- The simplistic nature of the proposed scores draws attention towards the pitfalls in the current state-of-the-art as well as the widely accepted definition of transferability scores.

- This paper highlights new research directions and postulates characteristics of an ideal transferability measure that will be helpful in streamlining future studies.

2. Background

2.1. Prior Art

Model-agnostic transferability of pre-trained models to a downstream task is a scarcely studied area in deep learning literature. Most of the existing methods have only been limited to evaluating the pre-trained models for classification tasks. In Nguyen et al. (2020b), Nguyen *et al.* proposed one of the first methods for quantifying the transferability of the pre-trained models for a given downstream task. Given the downstream training data, this method computes the predictions using the pre-trained model and estimates a joint distribution over the predicted labels and true labels to construct an empirical predictor. The log expectation of this predictor (LEEP) Nguyen et al. (2020b) is used as a measure for determining resonance between the pre-trained model and the downstream task. Building on this work, You et al. (2021) proposed to exploit the log of the maximum value of label evidence (LogME) given embedding extracted from pre-trained models as a measure of transferability. Unlike LEEP, LogME is generic in nature and can be used for both classification and regression tasks. Apart from these methods, some earlier studies Achille et al. (2019); Zamir et al. (2018) proposed the use of expensive optimisation of the pre-trained model over the downstream dataset to evaluate transferability. In another line of investigation, Suresh et al. (2023) proposed to exploit the topology of feature embedding space to quantify the transferability in a variety of neural architectures. Compared to LogME and LEEP, which require only one forward pass over the pre-trained models, existing studies are computationally very expensive to use in practice.

Comparison with the proposed method Both LEEP Nguyen et al. (2020b) and LogME You et al. (2021) quantify transferability of pre-trained models from a probabilistic standpoint. In contrast, the proposed method analyses the geometric structure of embedding space of a pre-trained model to quantify its transferability. Unlike the existing methods, the proposed scores don't require any complex distribution estimations. Moreover, the analysis of neural collapse-based geometric properties of embedding space is straightforward and computationally efficient. The proposed scores, as well as these existing methods, quantify transferability by analysing the discriminative nature of embedding/features generated by the pre-trained models for the downstream classification tasks. However, the proposed scores directly measure the discriminative characteristics of embedding while studying the geometric characteristics of the embedding space. On the other hand, both LEEP and LogME indirectly evaluate the discriminative nature of pre-trained embedding for downstream tasks.

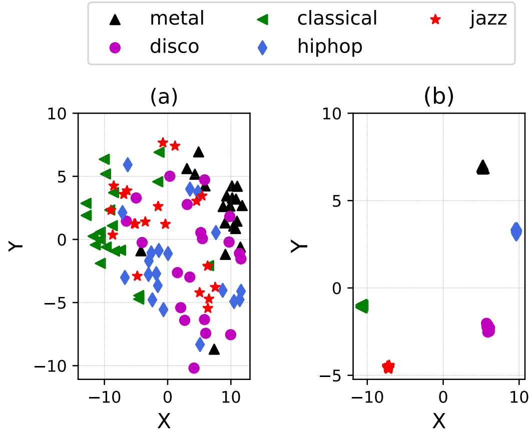


Figure 1: Penultimate layer embedding obtained from a convolutional recurrent neural network (CRNN) during (a) first and (b) last epoch of training for music genre classification. The last epoch exhibits variability collapse. All embeddings converge nearly to their class-specific embedding mean.

2.2. Neural collapse

Neural collapse defines the characteristics observed in the penultimate layer embedding, and the classification layer weights during the terminal stages of supervised training, and the same has been studied in the context of cross-entropy Papyan et al. (2020) and mean square error loss functions Han et al. (2021). These characteristics are listed below, and the authors encourage the reader to follow Papyan et al. (2020); Han et al. (2021) for more details:

- As the training progresses, within-class variances of penultimate layer embedding decrease. During terminal stages, these embeddings collapse to their respective class means. This behaviour is also known as *variability collapse*. Figure 1 illustrates the variability collapse observed during the training of a convolutional recurrent neural network (CRNN) for the task of music genre classification.

- The class mean embedding (centred with global embedding mean) converges to a simplex equiangular tight frame in the terminal stages.

- The re-scaled class mean embedding and linear classifiers (in the classification or last layer) converge in terminal stages, even though they lie in dual-vector spaces.

- The classification layer converges to the nearest neighbour classifier and assigns the class to an embedding whose class mean (of training data embedding) is nearest to it.

3. Neural Collapse Scores

3.1. Problem statement

Given a set of N pre-trained models $\{f_n\}_{n=1}^{N}$ and training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{K}$ for the downstream task, P_n represents the performance of f_n on downstream task (such as accuracy or area under ROC curve) after fine-tuning. We need to compute transferability score S_n for each f_n such that $\{S_n\}_{n=1}^{N}$ and $\{P_n\}_{n=1}^{N}$ correlate (either positively or negatively) with each other. Once we have obtained these transferability scores that are supposed to be positively correlated with the performance, an optimal pre-trained model $f_{n*} \in \{f_n\}_{n=1}^{N}$ can be selected as:

$$n* = \underset{n}{\operatorname{argmax}}\{S_n\}_{n=1}^{N} = \underset{n}{\operatorname{argmax}}\{P_n\}_{n=1}^{N}. \tag{1}$$

Here, argmax is replaced with argmin if transferability scores are supposed to be negatively correlated with model performance.

3.2. Neural collapse-inspired transferability scores

The proposed transferability scores are mainly derived from the variability collapse. For a given downstream task dataset $\mathcal{D}$ and a pre-trained model f_n, the transferability score S_n can be computed as:

$$S_n = \frac{\sum_{c=1}^{C} \left(\frac{1}{|\mathcal{D}_c|} \sum_{i=1}^{|\mathcal{D}_c|} ||\hat{f}_n(\mathbf{x}_i) - \boldsymbol{\mu}_c||_2 \right)}{\sum_{\forall (c,c')} ||\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}||_2}, \tag{2}$$

where $\hat{f}_n(\mathbf{x}_i)$ is the embedding obtained after penultimate layer, C is the number of classes, $\mathcal{D}_c$ and $\boldsymbol{\mu}_c$ are the set of examples & mean embedding belonging to class c, respectively.

The proposed transferability score S_n measures the degree of variability collapse observed in the penultimate layer embedding of pre-trained model f_n for the downstream training data $\mathcal{D}$ (without f_n being trained for the downstream task). As discussed earlier, a trained model is expected to exhibit variability collapse, i.e. all embedding of class c converge to class embedding means $\boldsymbol{\mu}_c$. Hence, the degree of variability collapse quantifies the deviation from this optimal structure, and the proposed score favours the pre-trained models that exhibit a lesser degree of variability collapse. Moreover, the proposed transferability score also penalises a pre-trained model if the average distance between class embedding means is much less. As a result, it favours the optimal pre-trained model that can generate semantically rich embedding for the downstream task. The proposed scores are always positive; a lesser score implies better transferability.

Although the variability collapse for the original dataset is observed only in the penultimate layers of the model, the same might occur at shallower depths for the downstream dataset. Hence, the proposed score is defined to be not restrictive to the embedding or features extracted from any specific model layer. Since we are conflating variability collapse to an embedding space, the proposed scores are generic and can be applied to any pre-trained model (supervised or unsupervised) or features.

Table 1: Characteristics of the pre-trained models used for COVID-19 prediction.

Model	# Parameters	Pre-trained Classification task
DCASE 2018 baseline	4.31 M	Acoustic scenes
CNN	0.43 M	Music genres Speech commands
CRNN	0.53 M	Music genres Speech commands

4. Experimental Details

4.1. Datasets

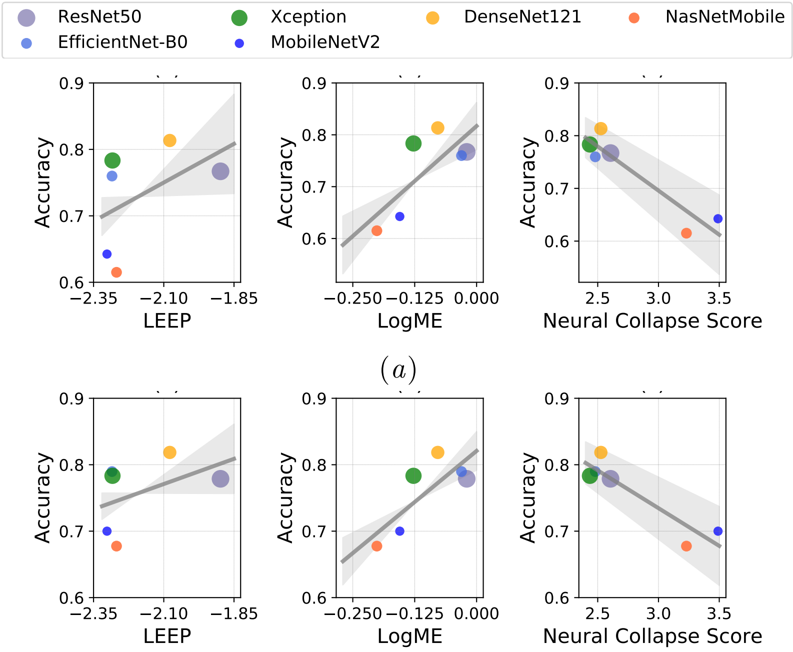
We evaluate the proposed scores for the downstream tasks of image classification using the CIFAR-10 dataset and cough sounds-based COVID-19 prediction using the COSWARA dataset Bhattacharya et al. (2023). We sample cough sounds from patients for whom the PCR test information was available as the meta-data. Overall, we selected 960 positive audio samples and 428 negative samples. The audio recordings are approximately $3 - 8$ seconds long, mono and sampled at 48kHz. Following Bhattacharya et al. (2023); we use a 64 log mel-bands Mel-spectrogram as input representation, extracted using a short-term Fourier transform (STFT) with an FFT window of 100ms with 50% overlap.

4.2. Pre-trained models and transfer learning

For CIFAR-10, we used ResnetNet-50, EffiecientNet-B0, Xception, DenseNet-121, NasNet-Mobile and MobileNet-V2 (trained on ImageNet dataset) as the pre-trained models mod (2022). On the other hand, Table 1 documents the pre-trained models used for COVID-19 detection. We have used a CNN baseline provided in Kong et al. (2018) for DCASE 2018 challenge dca (2018), a smaller version of this baseline model (*CNN*) with lesser layers and a CRNN that is derived from *CNN* by adding a GRU or recurrent layer after convolutional layers. *DCASE* architecture is pre-trained on DCASE 2018 acoustic scene classification (*DCASE-ASC*) dca (2018). *CRNN* and *CNN* are trained on speech command classification dataset Warden (2018) and GTZAN, a music genre classification dataset Tzanetakis et al. (2001). For acoustic scenes, music and speech signals, we have used 40 ms, 40 ms and 30 ms frames (50% overlap), respectively.

For transfer learning, we removed the last layer from the pre-trained models and added a linear classification layer for the downstream tasks. We train these models in two ways: training only the newly added classification layer while keeping the pre-trained layers frozen and training both newly added and the pre-trained layers. The performance trends of different pre-trained models on a downstream task are expected to be consistent across the transfer learning strategies You et al. (2021). Our implementation of neural collapse scores is publicly available[1].

[1] https://github.com/AnshThakur/NC_Scores_Demo



(a)

(b)

Figure 2: Relationship observed between transferability scores and model performance on CIFAR-10: (a) with training only the classification layer, the absolute Pearson coefficient between accuracy and LEEP, LogME, or neural collapse scores is 0.44, 0.76 and 0.93, respectively. (b) with fine-tuning the whole pretrained model, the absolute Pearson coefficient between accuracy and LEEP, LogME, or neural collapse scores is 0.47, 0.79 and 0.91, respectively.

4.3. Comparative methods and parameter setting

We compare the proposed transferability score against the LogME You et al. (2021) and LEEP Nguyen et al. (2020b) scores (see Section 2). We rank all the pre-trained models based on their transferability scores with respect to the downstream task. Note that, unlike neural collapse scores, the larger values of LEEP or LogME indicate better models.

The cough sounds dataset is divided into train (70%), test (15%) and validation (15%) sets. The number of images in training, validation and test sets is 40K, 10K and 10K, respectively, for the CIFAR-10 dataset. Each model is trained or fine-tuned for 250 epochs using a batch size of 32, cross-entropy loss and Adam optimiser with a learning rate set to 0.0001. Classification accuracy for the image classification task and area under the ROC curve (AUROC) for the COVID-19 detection task is used as a performance metric. Model checkpoints are used to find the best-performing model parameters on the validation sets. Note that the initialisation of the newly added classification layer is kept constant across all models for each comparative run. We perform each comparative experiment with 10 different initialisation seeds, and the average of 10 runs is presented here as the final performance.

5. Results and Discussion

Figure 2 illustrates relationships between the comparative transferability scores and the performance of different models on CIFAR-10 obtained by training only the classification head and re-training all layers, respectively. Similarly, Figure 3 reports the performance of different models for COVID-19 detection obtained by training only the classification head and re-training all layers, respectively. The analysis of these figures highlights the following:

- There is a *near perfect* correlation between the neural collapse scores of the pre-trained models and their performance after transfer learning on both CIFAR-10 and COVID-19 datasets. Based on the proposed scores, we can successfully pre-empt the better and the worst-performing models after transfer learning.

- Similar to the proposed scores, LEEP and LogME scores also exhibit good performance trends. These scores have been able to discriminate between the transferable and non-transferable models to a great extent. Although the absolute correlation between scores and model performance is better for the neural collapse, the performance of LogME is comparable across all experiments.

- Although *DenseNet-121* has been identified as one of the best-performing models on CIFAR-10 by all the scores, they were not able to identify *DenseNet-121* as the best-performing model.

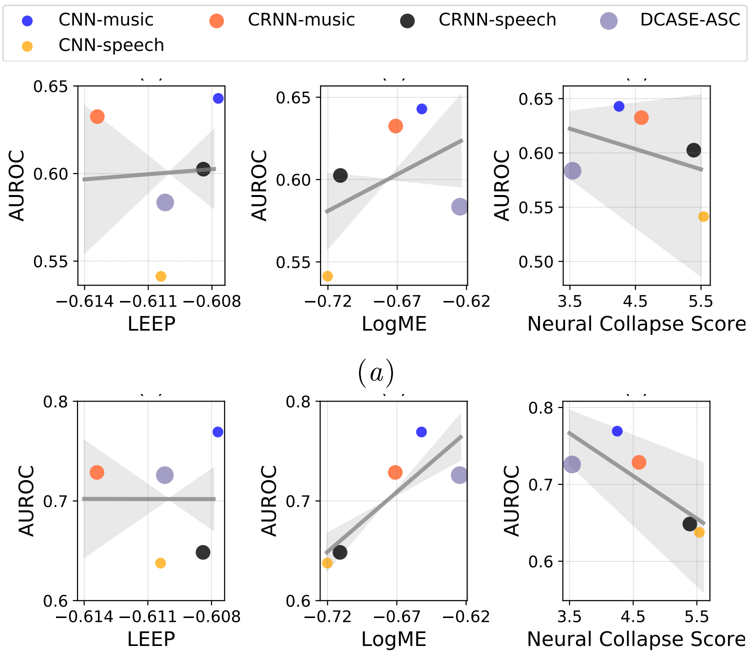
- Both LogME and the proposed scores outlined *DCASE-ASC* as the best performing model for COVID-19 detection. However, this model was outperformed by *CNN-music*, which was considered the second best by these scores. As mentioned earlier, despite these nuances, the proposed and existing scores have been able to capture performance trends with great success.

5.1. Caveats in neural transferability scores

The success of the proposed neural collapse scores and experimental results highlight some major drawbacks in the outlook being followed by current state-of-the-art in addressing neural transferability. These drawbacks are discussed below:

- Current neural transferability definition is mainly concerned with classification tasks. As a result, the discriminative nature of embedding space can be considered a valid transferability measure. One can directly measure this discriminative nature using simple variability analysis (as done by neural collapse scores). Hence, it is obvious to question the requirement of complex probabilistic estimations (as performed in the current state-of-the-art) to measure neural transferability.

- Transfer learning often surpasses the classification problems and has been used in a wider range of applications such as generative modelling, clustering and segmentation tasks. In such applications, the desired semantic meaning-fullness of embedding space goes beyond just a simple measure of class discrimination. Hence, the current transferability scores are not suitable for measuring generic neural transferability.



(*a*)

(*b*)

Figure 3: Relationship observed between transferability scores and model performance on COVID-19 detection: (a) with training only the classification layer, the absolute Pearson coefficient between accuracy and LEEP, LogME or neural collapse scores is 0.68, 0.84 and 0.86, respectively. (b) with fine-tuning the whole pretrained model, the absolute Pearson coefficient between accuracy and LEEP, LogME or neural collapse scores is 0.61, 0.83 and 0.85, respectively.

- All transferability scores are computed from the training data of the downstream task. A good transferability score on the training data may not resonate with the performance on the validation or test dataset. In case of training data scarcity or over-fitting, this behaviour is likely to be amplified. We witnessed this behaviour with *DCASE-ASC* model on the COVID-19 dataset. Despite being deemed the best model (see Figure 3) by LogME and neural collapse scores, the performance of *DCASE-ASC* was second to *CNN-music*. On analysing the training loss and validation scores obtained during fine-tuning of *DCASE-ASC* and *CNN-music* models (see Figure 4), note that *DCASE-ASC* model exhibited faster and better convergence. However, *DCASE-ASC* model failed to exhibit better generalisation on validation and test datasets. The lack of generalisation could be attributed to the scarcity of training data (only 971 training examples).

- Current transferability scores do not take into consideration the model size or complexity. In many applications, a smaller model would be preferred over a larger model if both of them provide similar scores.

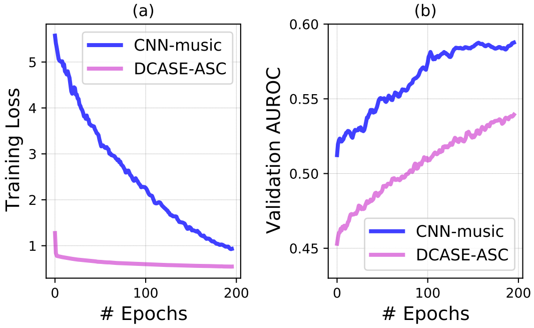


Figure 4: Training dynamics of *DCASE-ASC* and *CNN-music* models observed during fine-tuning for COVID-19 detection using cough sounds.

5.2. An ideal transferability measure

Based on the above-mentioned drawbacks, we can postulate an ideal transferability score that is characterised by the following properties:

- It should be generic in nature or agnostic to the type of downstream task.
- It should consider model complexity or the number of model parameters while determining the optimal model.
- It should exploit both training and validation data to get better estimates of generalisation for the downstream tasks.
- It should be significantly more computationally efficient than the trivial brute-force approach (fine-tuning all models and analysing their performance).

6. Conclusion

This work highlighted the major flaws in current neural transferability literature by equating state-of-the-art transferability measures to simple discrimination of embedding generated by the pre-trained models. As a tool to highlight these flaws, this work proposed neural collapse based transferability scores that exhibited either comparable or better performance in pre-empting the best-performing pre-trained model after transfer learning for classification tasks. The simplicity of the proposed scores highlighted that neural transferability for classification tasks is straightforward, and the community should strive for generic and practical transferability measures. We concluded this work by hypothesising the characteristics of an ideal transferability measure that could overcome most of the pitfalls of the current state-of-the-art. In future, we will be working towards the realisation of this hypothesised ideal transferability measure.

7. Acknowledgement

This work is supported by IIITD-IITD joint research grant (MFIRP-233) and Infosys Foundation via Infosys Centre for AI, IIIT Delhi.

References

Dcase 2018 challenge, 2018. URL http://dcase.community/challenge2018/index.

Keras Pre-trained Models. https://keras.io/api/applications/, 2022. Online; accessed 31st July 2022.

Jakob Abeßer. A review of deep learning based methods for acoustic scene classification. *Applied Sciences*, 10(6), 2020.

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439, 2019.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

Debarpan Bhattacharya, Neeraj Kumar Sharma, Debottam Dutta, Srikanth Raj Chetupalli, Pravin Mote, Sriram Ganapathy, C Chandrakiran, Sahiti Nori, KK Suhail, Sadhana Gonuguntla, et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection. *Scientific Data*, 10(1):397, 2023.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

Tomoya Koike, Kun Qian, Qiuqiang Kong, Mark D Plumbley, Björn W Schuller, and Yoshiharu Yamamoto. Audio for audio is better? an investigation on transfer learning models for heart sound classification. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 74–77, 2020.

Qiuqiang Kong, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D Plumbley. Dcase 2018 challenge surrey cross-task convolutional neural network baseline. *Parameters*, 4:4–691, 2018.

Cuong Nguyen, Thanh-Toan Do, and Gustavo Carneiro. Uncertainty in model-agnostic meta-learning using variational inference. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3090–3100, 2020a.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305, 2020b.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Jordi Pons, Joan Serrà, and Xavier Serra. Training neural audio classifiers with few data. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16–20. IEEE, 2019.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Sungho Shin, Jongwon Kim, Yeonguk Yu, Seongju Lee, and Kyoobin Lee. Self-supervised transfer learning from natural images for sound classification. *MDPI Applied Sciences*, 11(7), 2021. ISSN 2076-3417. doi: 10.3390/app11073043.

Suryaka Suresh, Bishshoy Das, Vinayak Abrol, and Sumantra Dutta Roy. On characterizing the evolution of embedding space of neural networks using algebraic topology, 2023. URL http://arxiv.org/abs/2311.04592.

Anshul Thakur, Daksh Thapar, Padmanabhan Rajan, and Aditya Nigam. Deep metric learning for bioacoustic classification: Overcoming training data scarcity using dynamic triplet loss. *The Journal of the Acoustical Society of America*, 146(1):534–547, 2019.

Anshul Thakur, Vinayak Abrol, Pulkit Sharma, Tingting Zhu, and David A. Clifton. Incremental trainable parameter selection in deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022. doi: 10.1109/TNNLS.2022.3210297.

George Tzanetakis, Georg Essl, and Perry Cook. Automatic musical genre classification of audio signals, 2001. URL http://ismir2001.ismir.net/pdf/tzanetakis.pdf.

Aäron Van Den Oord, Sander Dieleman, and Benjamin Schrauwen. Transfer learning by supervised pre-training for audio-based music classification. In *Conference of the International Society for Music Information Retrieval (ISMIR 2014)*, 2014.

Siddhant R. Viksit and Vinayak Abrol. Multi-layer acoustic & linguistic feature fusion for ComParE-23 emotion and requests challenge. In *ACM International Conference on Multimedia*, page 9492–9495, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612851.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

P. Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv e-prints*, April 2018. URL https://arxiv.org/abs/1804.03209.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pages 12133–12143, 2021.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

Hengshun Zhou, Xue Bai, and Jun Du. An investigation of transfer learning mechanism for acoustic scene classification. In *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 404–408, 2018. doi: 10.1109/ISCSLP.2018.8706712.