Model Transferability Informed by Embedding's Topology

Felipe Gutierrez FIGUTIER@UC.CL
Hans Lobel HALOBEL@UC.CL
Faculty of Engineering, Pontificia Universidad Católica de Chile

Abstract

In this work, we tackle the challenge of predicting the performance of a pre-trained classification model on a downstream task before fine-tuning. Our approach leverages the geometric information encoded in the feature embeddings of pre-trained networks, which we analyze using persistent diagrams generated from a Vietoris-Rips filtration. We find that during late-stage training, the separation between the highest-persistence features and the remaining low-persistence features mirrors the dynamics of neural collapse. However, our topological measures differ significantly during early training as the geometrical structure of the embeddings stabilizes. We propose a transferability score based on the ratio of these topological features. We evaluated its performance in ranking models for fine-tuning and showed that it achieves competitive results against established methods.

Keywords: Topological Data Analysis, Persistent Homology, Model Selection, Representation Learning, Fine-Tuning, Neural Collapse

1. Introduction

Selecting the optimal pre-trained model for a new task is a critical challenge, as exhaustively fine-tuning every option is computationally prohibitive. Transferability estimation (You et al., 2021; Tran et al., 2019; Nguyen et al., 2020) addresses this by creating efficient heuristics to rank models without requiring full fine-tuning.

Recent work has established a connection between feature geometry, model performance, and transferability, specifically highlighting the role of neural collapse (Ding et al., 2023; Wang et al., 2023; Suresh et al., 2023). Building on this evidence, our approach is guided by the intuition that a highly transferable model should produce representations where features of the same class are compact and features from different classes are well separated. We propose using persistent homology (Edelsbrunner et al., 2008; Edelsbrunner et al., 2002) to analyze the connectivity of embeddings across multiple scales, offering a global and robust perspective on the representation's geometry.

We base our analysis on persistence diagrams obtained using Vietoris-Rips filtration on the embedding space generated by the models, which naturally decompose features by their topological structural importance. High-persistence features correspond to large, stable structures like well-separated class clusters, whereas low-persistence features capture local noise and finer-grained connections.

We take advantage of this natural decomposition of topological characteristics to define novel measures that quantify desirable geometric properties, allowing us to track the evolution of the separation of the clusters and the tightness within the cluster and evaluate the suitability of a model for transfer learning.

Experiments show that the proposed measures effectively track the progression of features during training, and, more importantly, reach state-of-the-art performance when ranking pre-trained models for fine-tuning tasks.

2. Topological Measures for Representation Analysis

By processing the 0-dimensional persistence diagram generated from a Vietoris-Rips filtration of a network's embeddings, we can quantify topological structures that describe the degree of separation between clusters and the degree of tightness within those clusters. For a classification problem with k classes, a representation is expected to form k tight, well-separated clusters (Papyan et al., 2020). This manifests itself as a distinct signature in the persistence diagram, as illustrated in Figure 1. Specifically, the highest persistence values $k-1$ are expected to be high due to the length of the radius that the connected components need to merge between samples of different classes in the context of Vietoris-Rips filtration. In contrast, the remaining $n-(k-1)$ values should be close to zero due to the fact that the radius required by these samples to merge is small, reflecting the tightness converging to the mean of the samples toward their respective class means.

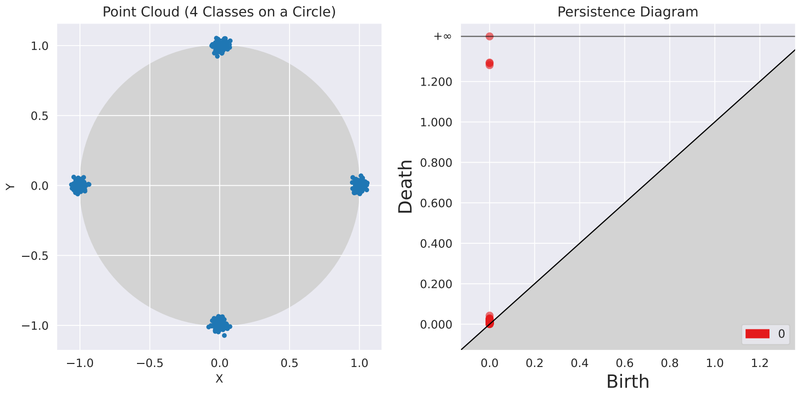

Figure 1: Persistent diagram of 4 equidistant classes sampled with low variance in a circle. A representation with high inter-cluster distance and low within-cluster variability should have high persistence on the $k-1$ on the most persistent modules and near 0 on the remaining ones.

Let the persistence diagram D_0 contain n off-diagonal points. Let p_i denote the i-th persistence value $(d_i - b_i)$, sorted in descending order. We propose the following measures:

$$\mathcal{H}_{\text{sep}} = \frac{1}{k-1}\sum_{i=1}^{k-1} p_i, \qquad \mathcal{H}_{\text{tight}} = \frac{1}{n-(k-1)}\sum_{i=k}^{n} p_i, \qquad \mathcal{H}_{\text{ratio}} = \frac{\mathcal{H}_{\text{tight}}}{\mathcal{H}_{\text{sep}}} \tag{1}$$

In the late stage of training, $\mathcal{H}_{\text{sep}}$ captures the separation between the main class clusters, while $\mathcal{H}_{\text{tight}}$ measures the compactness of points within each cluster. Finally, the ratio $\mathcal{H}_{\text{ratio}}$ quantifies the importance of intra-cluster tightness relative to the inter-cluster distance.

3. Experiments and Results

We empirically evaluate our proposed topological measures by first assessing whether they reflect the expected dynamics after the late stages of training, *i.e.*, high separation between classes, low variance within classes. Next, we validate their ability to rank pretrained models' fine-tuning performance by benchmarking them against a series of established methods.

3.1. Evolution of Topological Measures

We trained a ResNet-18 on the CIFAR-100 dataset. We computed the metrics on random subsamples of the training set embeddings, using 2,048 samples for the PD computation.

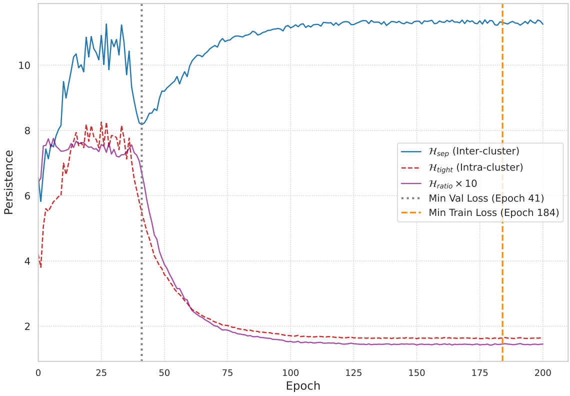

Figure 2: Evolution of the proposed metrics on ResNet18 trained on CIFAR-100

As can be seen in Figure 2, our topological measures, $\mathcal{H}_{\text{sep}}$ and $\mathcal{H}_{\text{tight}}$, effectively track the expected dynamics. During training in the late stages, the separability between the groups $\mathcal{H}_{\text{sep}}$ presents stable and high values, while the tightness between groups $\mathcal{H}_{\text{tight}}$ decreases. Meanwhile, $\mathcal{H}_{\text{ratio}}$ exhibits a decreasing pattern as soon as the network begins to converge and stabilizes in the late training stage. This behavior validates that our measures capture the expected geometric changes of a maturing representation, where class clusters become well-separated and points within them become more compact. The measures are more erratic during early training, a behavior we attribute to the global sensitivity of our Vietoris-Rips filtration method to outlier points before the class clusters have fully formed.

3.2. Predicting Fine-Tuning Performance

For our transferability benchmark, we evaluated a diverse set of pre-trained models from the PyTorch model zoo, replicating the datasets and models used on You et al. (2021); all models are pre-trained on ImageNet-1k (Deng et al., 2009).

To benchmark transferability, we computed $\mathcal{H}_{\text{ratio}}$ for the embeddings of each pre-trained model on the source dataset, ImageNet-1k, and on the target datasets. Then, we ranked the models based on these scores and evaluated the quality of this ranking by comparing it to the models' performance after fine-tuning. To this end, we use Kendall's τ (Kendall, 1938). The results of this procedure are summarized in Table 1.

Table 1: Kendall's τ Correlation with Target Accuracy. For each dataset (row), the highest correlation is in **bold** and the second-highest is underlined. PR stands for $-\mathcal{H}_{\text{ratio}}$ We compare it against LEEP (Nguyen et al., 2020), NCE (Tran et al., 2019), LogME (You et al., 2021), and the NC Score (Suresh et al., 2023).

Dataset	**LEEP**	**NCE**	**LogME**	**NC Score**	**PR**	**PR (ImageNet)**
Aircraft	0.289	0.511	0.360	**0.644**	**0.644**	0.556
Birdsnap	0.333	0.644	0.467	**0.778**	0.733	0.689
Caltech101	0.422	**0.733**	0.584	0.422	**0.733**	0.689
StanfordCars	0.424	0.566	0.519	**0.801**	0.283	0.660
CIFAR10	0.477	0.159	**0.705**	0.205	0.341	0.614
CIFAR100	0.422	0.111	0.629	0.244	−0.067	**0.644**
DTD	0.022	−0.289	0.494	0.200	−0.556	**0.689**
Pets	0.467	**0.733**	0.511	0.556	0.378	0.556
SUN397	0.449	0.764	0.719	0.315	0.405	**0.809**
Mean	0.367	0.437	0.554	0.463	0.322	**0.656**

We find that the negative ($-\mathcal{H}_{\text{ratio}}$) extracted from the original pre-training dataset features has a consistent performance and outperforms every other method on average, suggesting that the intrinsic topological structure of a source model's representation is a powerful indicator of its transferability. In contrast, when applying the negative Persistence Ratio $-\mathcal{H}_{ratio}$ to the target task embeddings, the results were inconsistent. Although it achieved strong performance on specific fine-grained datasets like Aircraft and Birdsnap, its overall reliability was lower than established methods.

Like $-\mathcal{H}_{ratio}$ on the target task, the score proposed by Suresh et al. (2023) shows high performance on some datasets, but very low performance on others. $\mathcal{H}_{ratio}$ exhibits an extreme version of this instability, with a strong negative correlation on DTD. However, in some datasets where the metric aligns with performance, the results are more competitive than in all other tested methods. This suggests that the topological signal is exceptionally strong but sensitive to the specific features of the target task.

4. Conclusion

We proposed topological measures that quantify the geometry of neural embeddings, which aim to describe the dynamics in which clusters get defined across training. We find that the intrinsic topological structure of a source model's representation is a powerful and reliable predictor of fine-tuning performance. Applying these measures to the target task revealed a more complex relationship, suggesting that future work should explore the geometric alignment between source and target tasks. Ultimately, our work presents evidence of how the embedding has information related to the performance and transferability of a model.

Acknowledgments

The authors acknowledge the support of the National Agency for Research and Development ANID-Subdirección de Capital Humano/Doctorado Nacional/2024-21241334, the National Center of Artificial Intelligence (ANID grant FB210017), the Millennium Institute for Foundational Research on Data (ANID grant ICN17_002), and FONDECYT grant 1241772.

References

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Yuhe Ding, Bo Jiang, Lijun Sheng, Aihua Zheng, and Jian Liang. Unleashing the power of neural collapse for transferability estimation, 2023. URL https://arxiv.org/abs/2310.05754.

Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002. ISSN 1432-0444. doi: 10.1007/s00454-002-2885-2. URL https://doi.org/10.1007/s00454-002-2885-2.

Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453(26):257–282, 2008.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7294–7305. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/nguyen20b.html.

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.

Suryaka Suresh, Vinayak Abrol, and Anshul Thakur. Pitfalls in measuring neural transferability. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023. URL https://openreview.net/forum?id=KWUA0n6Dpv.

Anh T. Tran, Cuong V. Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5549–5558, October 2023.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12133–12143. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Appendix A. Training Hyperparameters

All models were trained from scratch for 200 epochs using a batch size of 128. We used a Stochastic Gradient Descent (SGD) optimizer with a standard momentum and weight decay configuration. The loss was computed using Cross-Entropy Loss. To adapt the learning rate, we employed a ”reduce learning rate on plateau” scheduler that monitored the training loss. The scheduler would reduce the learning rate by a factor of 0.5 if the loss did not improve for a patience of 5 epochs.

The specific hyperparameters that varied between experimental runs, such as the initial learning rate and metric-specific sample sizes, are detailed in Table 2.

Table 2: Hyperparameter Configurations for Model Training.

Architecture	**Dataset**	**Initial LR**	**PD Sample Size**	**NC Sample Size**
ResNet-18	CIFAR100	0.0679	2048	10000
DenseNet-121	CIFAR100	0.0679	2048	10000
ResNet-18	FashionMNIST	0.0100	512	5000
DenseNet-121	FashionMNIST	0.0100	512	5000
ResNet-18	MNIST	0.0100	512	5000
DenseNet-121	MNIST	0.0100	512	5000

Appendix B. Additional Correlation Analyses

To provide a more comprehensive evaluation of the transferability metrics, this appendix presents supplementary correlation results. These tables are intended to give additional perspective on the Kendall’s τ values reported in the main body of the paper. It is important to note that the evaluated models and the ground-truth ranking for the benchmark are derived from the benchmarking and fine-tuning performance results reported in (You et al., 2021)

It is important to note that Kendall’s τ values are typically lower in magnitude than other correlation coefficients like Pearson’s. A value that might seem moderate, such as 0.4, can indicate a strong and meaningful association in ranking performance.

To offer a more intuitive interpretation, we include Table 3, which shows the Concordant Probability, also known as Pairwise Accuracy (P_c). This value is derived from Kendall’s

τ via the formula $P_c = (\tau + 1)/2$ and represents the probability that a given metric will correctly order a random pair of models according to their final fine-tuning accuracy. A value of 0.5 corresponds to random chance. This translation helps contextualize the Kendall's τ values; for example, a τ of 0.4 corresponds to a pairwise accuracy of 70

Finally, we also provide the Spearman's ρ correlation coefficients in Table 4 to demonstrate that the ranking trends are robust and not an artifact of a single statistical test. These additional results reinforce the conclusions presented in the main text.

Table 3: Pairwise Accuracy of Transferability Metrics Across Tasks.

Dataset	**LEEP**	**NCE**	**LogME**	**NC Score**	**PR**	**PR (ImageNet)**
Aircraft	0.644	0.756	0.680	**0.822**	**0.822**	0.778
Birdsnap	0.667	0.822	0.733	**0.889**	0.867	0.844
Caltech101	0.711	**0.867**	0.792	0.711	**0.867**	0.844
StanfordCars	0.712	0.783	0.759	**0.901**	0.641	0.830
CIFAR10	0.739	0.580	**0.852**	0.602	0.670	0.807
CIFAR100	0.711	0.556	0.815	0.622	0.467	**0.822**
DTD	0.511	0.356	0.747	0.600	0.222	**0.844**
Pets	0.733	**0.867**	0.756	0.778	0.689	0.778
SUN397	0.725	0.882	0.860	0.657	0.702	**0.905**
Mean	0.684	0.719	0.777	0.731	0.661	**0.828**

Table 4: Spearman's ρ Correlation with Target Accuracy.

Dataset	**LEEP**	**NCE**	**LogME**	**NC Score**	**PR**	**PR (ImageNet)**
Aircraft	0.467	0.697	0.462	**0.782**	0.770	0.721
Birdsnap	0.467	0.782	0.648	**0.915**	0.879	0.842
Caltech101	0.527	0.842	0.675	0.552	0.842	**0.855**
StanfordCars	0.562	0.704	0.642	**0.864**	0.401	0.778
CIFAR10	0.591	0.274	**0.811**	0.341	0.427	0.774
CIFAR100	0.564	0.273	0.742	0.394	−0.127	**0.818**
DTD	0.067	−0.345	0.693	0.236	−0.661	**0.842**
Pets	0.721	**0.879**	0.733	0.685	0.697	0.758
SUN397	0.535	0.815	0.821	0.395	0.523	**0.936**
Mean	0.500	0.547	0.692	0.574	0.417	**0.814**