International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Persistent Homology for Structural and Overfitting Detection in Transformer Models

Pranali D. Bhaisare[1], Dr. S. B. Kishor[2]

[1]Department of Mathematics, Rashtrapita Mahatma Gandhi College,
Nagbhid. Dist- Chandrapur
Emai ID: rutujanihar87@gmail.com
[2]Associate Professor, Department of Computer Science, Sardar Patel Mahavidyalaya, Chandrapur

Abstract

Transformer models are basic and important to modern systems of natural language processing. They provide high dimensional embeddings that learn meaningful relations in the process of training. Traditional evaluation metrics like the training-validation loss gap, can find the problem of overfitting; but these embeddings avoid the evolving geometric structure of it. In this study, we used topological tool, a Persistent Homology to dissect the topological characteristics of Transformer embeddings throughout training. In this application we can constructing filtrations from embedding distances for track the birth and persistence of features such as connected components (0D) and loops (1D) These are demonstrated through the use of persistence diagrams and Betti curves. The analysis shows that overfitting leads to short-lived, unstable topologies and this is an indicator of excessive adaptiveness to noise using the one hand on the other hand, well-generalised models have stable and long-lived topological structures. This model-agnostic technique can detect overfitting, and offers multi-scale insights to the learning process, and is also a complementary technique to traditional diagnostic metrics. Accordingly, Topological Data Analysis is a newly emerging field that holds great potential to improve the interpretability and robustness of Deep Learning Architectures.

Key words: Topology, Persistent, Topological Data Analysis, Homology, Betti curve, Overfitting

1. Introduction

The coming of Transformer architectures has started paradigm shift in natural language processing. In 2017, Vaswani, Shazeer, et al. came out with "Attention Is All You Need" and so the Transformer revolution began in earnest. [1] Unlike traditional recurrent neural networks, Transformers use of a self-attention mechanism to process sequential data in parallel and thus its simplifies the capture of long-range dependencies and generates embeddings with high semantic and contextual information. Based on this architecture, modern models like BERT and GPT have been found to be necessary tools in various Natural Language Processing tasks. [2]

The opacity of deep learning systems is a major challenge for interpretable and robust systems, specifically overfitting. Overfitting is when generalising to new data will lead to loss of model generalisation, as the model will have internalised idiosyncratic noise in the training data. This is often indicated in Transformers by an expanding difference between the training and validation loss. [3]

Nonetheless, conventional evaluation strategies are not wont to take into account the dynamical geometrical configuration of embeddings. These embeddings has high dimensionality space that represented as a point cloud spaced and the encoded learning trajectory of the model through their spatial organisation. The topological structure of such point clouds is an encoding of salient information regarding dynamics of learning. Topological Data Analysis (TDA) and Persistent Homology (PH) offer a powerful structure in which to explore the multiscale structure of such data. Through the analytical lens it is possible to separate meaningful topological features (such as clusters and loops) from transitory and noisy artefacts.

In this paper, persistent homology use for the structural analysis of transformers embeddings, and in the detection of over-fitting. A Vietoris-Rips filtration based on pairwise embedding distances has been produced. [4] The birth and death points of zero-dimension connected components and are one-dimensional loops are evaluated by persistence diagram and betti curves. The analysis shows that the more intense the over-fitting, the more short-lived topological features are found to be, thus indicative of internal increased complexity. On the contrary, well-generalised models are stable topological features. This model-agnostic approach has made it possible to detect overfitting and offers multi-scale understanding of learning, and also complements traditional diagnostic metrics.

The major contribution of this research is as follows:

1. A new topological-data analysis pipeline has been proposed towards the real-time-monitoring of transformer embeddings.

2. Empirical evidence from persistence diagrams and Betti curves shows an explosion in short lived topological features related to overfitting.

3. Persistence diagrams and Betti curves are visualization tools for providing interpretable diagnostics of the topological structures that are contained in Transformers embeddings.

2. Background

2.1 Transformer Models

Transformer models are a class of deep learning architecture that has designed to process sequential data by using an attention mechanism. Transformers process input tokens using parallelism which enhances the computational efficiency and enabling the nuanced modelling the long-range interrelation within sequences. it is apposite to traditional sequence models such as recurrent neural networks (RNNs). The main part of the model is the self-attention module, which enables the model to determine the relative importance of the tokens in a sequence. [5]

This paradigm provides room for a full understanding of contextual relations that transcend entire input sequences. During training, Transformers adhere to generate high dimensional vector representations (embeddings) for every input token. These embeddings got semantic and contextual information which has been learned from the training corpus. As learning continues, these depictions of the embeddings vary from epoch to epoch, forming a rather complex geometrical structure in high dimensional space. Contemporary language models such as BERT are instantiated from Transformer architecture that generates contextualised embeddings which have been widely used in various applications of Natural

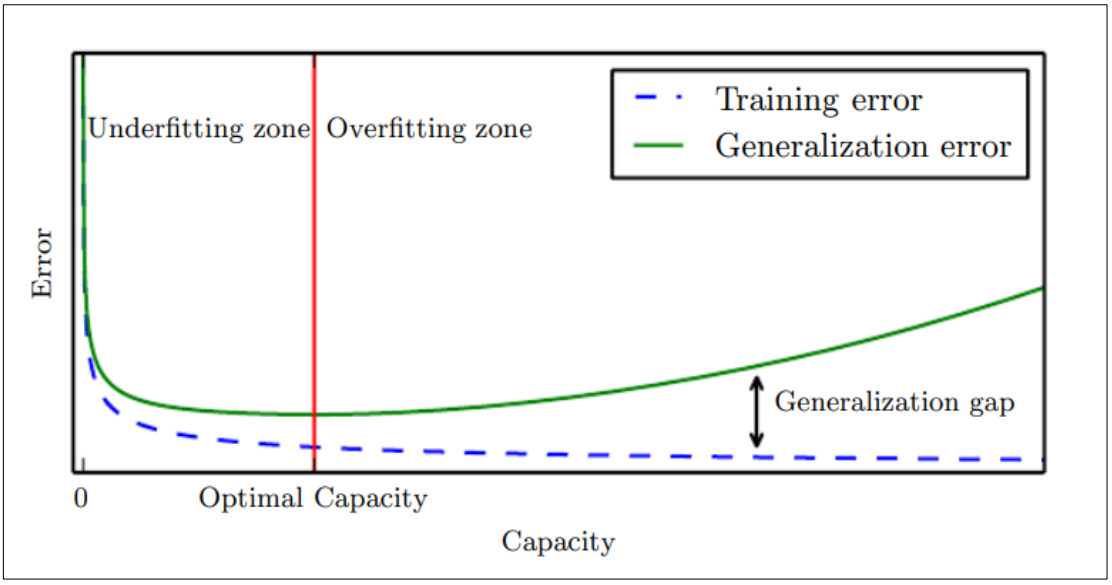
Language Processing (NLP).[6] These embeddings meaningful abstractions of textual data, and can be evaluable to gain insight into the nature of what is happening inside the model.

From a topological perspective, the embedding space generated by Transformers can be thought of as a collection of points in a high dimensional space. The space consumption of these points reflects the way the model lumps together or differentiates different semantic patterns within data. Investigating geometrical and topological properties of these embeddings provides interesting information about the dynamics of learning of the model and can help detect phenomena like over- fitting. This is research using topological approaches to examine embedding representations produced by Transformer models and allow exploration of structural patterns during training, and the identification of possible over-fitting.

2.2 What Is Overfitting in Deep Learning?

Overfitting is a simple and prevalent issue on machine learning and deep learning models. It emerges when that model assumes idiosyncratic patterns/ noise existed in the training data or massively failed to perform on new data or the unsee data. In such cases, the model will show good performance in the training data, whereas poor performance in the validation or a test data.

In the case of deep neural networks, overfitting is largely a result of the sheer number of parameters that there are, and the high capacity that follows this. In the process of optimisation the network can accidentally incorporate redundant noise/superfluous details that will cause the training to keep getting lower and the validation to keep higher or remain the same. [7]



Source: I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016. Chapter – 5, Pg no. 115

In Transformer based models, overfitting takes another form, which can be seen in the structure of the learned embedding as well. These models are generating high dimensional things as embeddings for each input token. If the system ends up memorising the training data with too much precision, then the embedding space inevitably becomes inordinately intricate and very hard to interpret, defeating the

purpose for the model to generalise. [8]. This is the point of concern and has been supported by latest empirical observations

Traditional diagnosis methods of overfitting, such as the increasing gap between training and validation loss, validation accuracy dropping, increasing model complexity, etc. give a surface-level answer. They do not reveal more fundamental changes taking place in the internal structure of the network representation. Therefore, more analytical tools are needed for the investigation of the structural behavior of neural network representations. In this context, this study takes the approach of topological techniques, the most notable Persistent Homology, to topological framework of Transformer embeddings, that provide an auxiliary approach so as to overfitting can be detected during the training process.

2.3 Topological Data Analysis (TDA)

Topological Data Analysis has become a powerful analytical tool that is gradually increasing in popularity as one method of interrogating the geometric shape and structural fabric of complex, high dimensional datasets. Whereas conventional statistical methods are mostly focused on the number of quantities and the associative relationship between them, TDA tries to capture the very plum of data through characteristics of its topology - its connexion, holes and higher dimensional voids as per the foundational works [9]. If we can study the topological structure of the data, we can detect the patterns, clusters, and the structural features of the data that are not easily reflected by the traditional analysis methods.

At the core of TDA is the realization that the geometrical structure of a dataset in space encodes some information about the inner organisation of the dataset. In many applied scenarios data lives in such high dimensionality that either we cannot visualize it or apply traditional analytics techniques to the data. TDA overcomes this limitation by asking questions about the properties such as connectivity, loops and higher dimensional cavities in order to reveal underlying structure relationships between data points [10]. Among a whole host of TDA tools, the most popular algorithm for this task is Persistent Homology. It captures birth and death of topological features at a continuum of spatial scales, and will thus allow for discrimination of important structural motifs from meaningless noise.

In deep learning, TDA is also increasingly employed in studying the internal observations within the neural networks. The high-dimensional embeddings output by neural networks may be treated as a point cloud lying in Euclidean space [11]. Applying topological techniques to such embeddings is enabling the exploration of the structure quality of the learned representations and give important insights into the process of learning. In the present work, we exploit Persistent Homology to examine the structural characteristics of Transformer derivable embeddings during the training process, to try and identify early periods of overfitting encoded in the topological shape change.

3 Persistent Homology

Persistent Homology is one of the pillars of Topological Data Analysis and is focused on identifying structural features of data at different spatial scales. It is a systematic identification and tracking of topological features, such as connected components, loops and voids, by using a filtration that scales the underlying metric [12]. A problem of analyzing the persistence of such features is distinguishing important structural patterns from irrelevant noise. In machine learning methods, data points are frequently in relation to the high-dimensional representations in embeddings that are generated by neural networks.

These embeddings can be perceived as points in a geometric space, in which the relationships in between data points is represented by the structure learned by the model. Persistent Homology challenges this structure by incrementally increasing a distance threshold and thereby observing the materialisation called connection among data points, their merging or dissipating [13].

When the distance threshold expands proximate points are connected and form a simplicial complex which represent geometric architecture data In the beginning, all the points are separate connected components. With the increasing scale, these components coalesce and higher dimension structures like loops or voids may appear [14]. Over the course of the filtration, Persistent Homology measures the time of life of each topological feature, [15] which leads to a computation of its life span, which is a measure of persistence. Features with long lifespans are conventionally assumed to have a lot of structural significance whereas fleeting features are discarded as being artefacts of noise.

The Topological features were categorised by the dimension. Zero dimensional (0D), features are equivalent to connected components and effectively summarising clusters of data. One-(1D) features are loop or cycles which shows circular relationship in the data points. Features of higher dimension are voids or cavities represent more intricate void a structures in the data [16]. Through the systematic analyzing of these topological features, Persistent Homology gives a powerful tool for study of the geometry structure of high - dimensional datasets. In our current investigation we have used Persistent Homology on Transformer-based embedding representations and in so doing, we explore the presence of structural motifs that indicate potential overfitting in training.

4. Structural Analysis of Transformer Representations

Transformer models generate high dimension vectorial representations, i.e. embeddings, of every constituent token. [17] These embeddings gather meaningful and related information learnt during training process. The way that these embeddings are organised in the vector space provides a glimpse into the internal relational schema of the model, providing important information on its learning process. In this study, we explored the Transformer embeddings from a topological perspective. Each embedding is considered as a point in a space with a high number of dimensions [18] i.e. Euclidean space. Pairwise distances between these points encode the relationships between similarities that the model has internalized throughout the model training. To bring out the structural arrangement of these points, we used the standard TDA methodologies. Specifically, we created a distance matrix using the embeddings, and then created a simplicial complex that captures their connectivity structure with different distance thresholds [19].

As distances threshold are increased, the nearby embedding points are started to connect and starting to form cluster and higher dimensional. This process leads to some filtration by which we can identify topological features such as connected components and loops. [20] By analyzing the birth and death of these features at the different scales it is possible to study the structural properties of the learned representation space. The analysis of these topological structures seems to bring important information on the organization of the embedding space created by the transformer models. Stable topological patterns provide meaningful groupings but unstable structures can be noisy and complex representation. Understanding these structural features is important to understanding the dynamics of learning of the neural networks. In particular, changing topologies of the embeddings can offer valuable indicators to

detect phenomena like overfitting during training. One-dimensional (1D) features, such as loops or cycles, are cyclic relationships in a data. In comparison, higher dimensional features capture voids or cavities in the data structure. [21] As a result, by analysing these topological features, Persistent Homology provides us with a powerful tool to understand the geometrical structure of high dimensional data sets.[22] This is using Persistent Homology helps represent what transformer models learn, to scared to observe structural patterns and spot potentially overtraining them during the training process.

5. Topological indicators of Overfitting

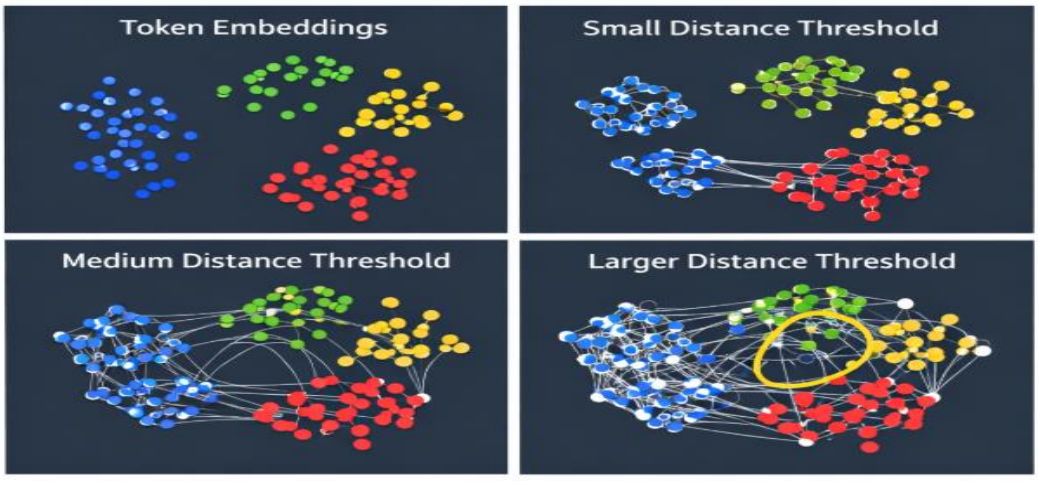
Overfitting occurs when the model learns specific patterns out of the training data as oppose to the general applicable and generalisable features. In such cases, traditional indicators, such as training and validation losses provide numerical evidence for overfitting, but do not provide any information about the structural behaviour of learned representations.

Topological Data Analysis (TDA) is another means of thinking about the structural properties of data. By applying Persistent Homology to the embedding representations that come out of a transformer model, it is possible to understand the evolution of the topological structure of the data during training. one can analyze how the topological structure of the data evolves during the training process.

When an overfitting occurs, the embedding space would usually become more structural in complex. This leads to unstable or short -lived topological features. These kinds of features are usually credited to small clumps or loops being in existence for only a fine range of filtration scales. In contrast, models which were good generalizers were often with more stable and long-lived topological structure. Therefore, rapidly changing/s short lived topological features can be used as an indicator of overfitting. By tacking these changes in architecture over the various training process, topological analysis gives more insights to learn the dynamics of transformer models.

6. Visualization with Persistence Diagrams and Betti Curves

Results obtained from Persistent Homology are visualized in most cases with persistence diagrams and Betti curves. These techniques provide an intuitive representation for the topological structure in high dimensional data. Such visualization techniques make it possible to see the characteristic changes of the topological features on different scales, as well as the meanings distinguishing the pattern in the structure from noise. [23]



In the figure, the token embeddings in high dimensional space on upper left, illustrating that each coloured point represents different semantic groups(clusters), these clusters are clearly separated, that show a well-defined and meaningful structure. This usually represents a well generalized model. The top-right panel, depicting a "small distance threshold", shows that points start to connect to their closest neighbours, forming small local clusters; specifically, edges begin to appear in the clusters and local structure preserved. The model is learning explicit relationships.

At the medium distance threshold, in the lower left, we observe that more edges start to appear, distinct clusters start to be partially interconnected, and some loops start to form; correspondingly, 1D topological features start to form. This creates a balanced structure encompassing both local and global relationships, often representing an ideal stage for learning. At the large distance threshold, located in the bottom-right corner, almost all points are interconnected, and abundance of edges creates a dense network. the highlighted part shows a loop structure; the overall structure becomes hyper-connected, and distinct clusters lose their identity signifying that overfitting has occurred.

Conceptual illustration of the evolution of topological structures in transformer embeddings across increasing distance thresholds, highlighting the emergence of loops as an indicator of overfitting.

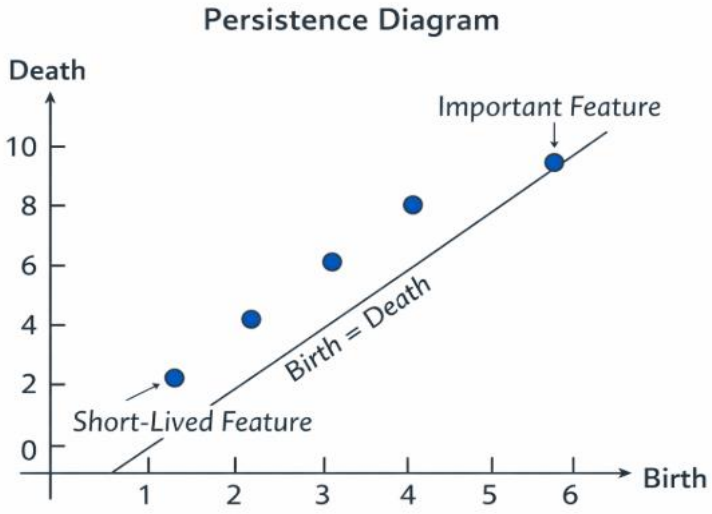


Figure 1.

In a persistence diagram the birth and death of topological features that are determined during the filtration process is represented. Each point in the diagram is based on a topological feature, where the horizontal axis (x-axis) is the scale of birth and the vertical axis (y-axis) is the scale of death. Features with greater persistence than ranges on the larger scale are farther away from the diagonal, and generally considered to be more significant. In contrast, points close to the diagonal are features of short duration, and are often attributes of noise. If death value and birth value of feature is nearly equal that feature does not exist long time Therefore, such short-lived features are considered as noise. Overfitting diagnosis can be done with persistent homology by studying the distribution of topological features in the persistence diagram. When a model starts over fitting, more short-lived features start appearing near the diagonal and this is an

indication that the model is starting to learn a noise, rather than something interesting. In contrast, non-well-generalized models have more stable (persistent) features in a further out region from the diagonal. Thus, when there are too many features which are not highly persistent, this is a hint from topology of this model that it is overfitted.

Beyond persistence diagrams, Betti curves can be another interesting way of representing topological information. Betti curves are descriptions of the number of topological features present on different scales of filtration. For example, the minimum of the scale parameter, the Betti-0 curve, is the number of connected components while the Betti-1 curve is the number of loops extant in the data structure as the sc

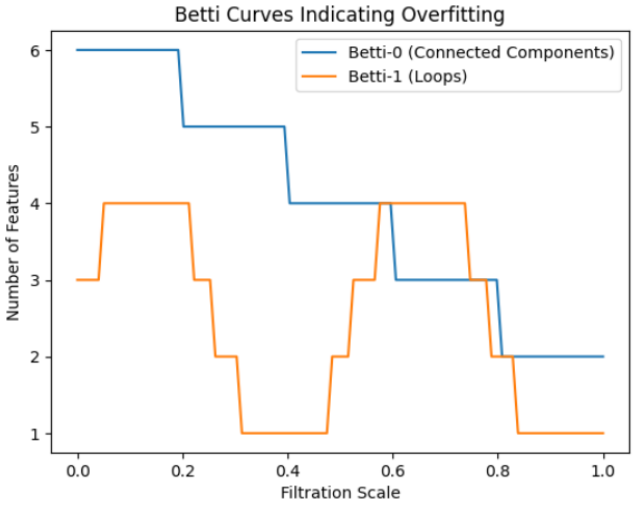


Figure 2.

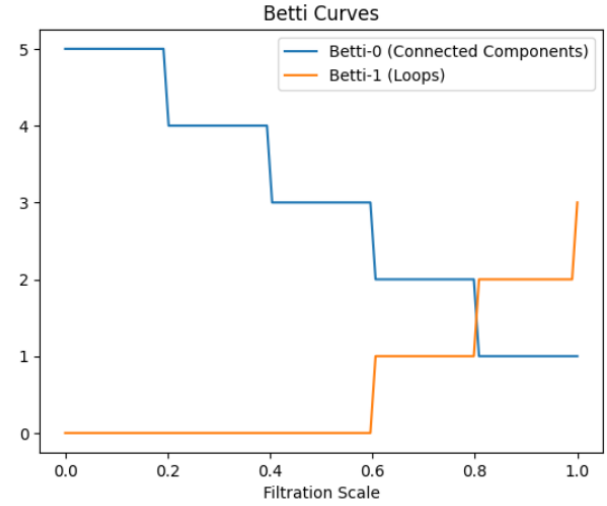


Figure 3.

In figure 3. The Betti curves indicate a stable topological structure of the embedding space. The smooth decrease in Betti-0 values reflects coherent clustering, while the moderate increase in Betti-1 values suggests meaningful structural relationships without excessive complexity. The absence of highly fluctuating or dominant one-dimensional features indicates that the model does not exhibit significant overfitting. In Figure 2. Betti curves exhibit symptoms of overfitting, including irregular changes in Betti-0 and large fluctuations in Betti-1 values. The presence of numerous unstable one-dimensional features suggests that the model is capturing noise instead of learning generalized structural patterns and is creating overly complex representations.

By analysing these curves, it is possible to gain an understanding of the changes in complexity of the structural of embedding space during the training process. Stable patterns as detected in Betti curves can indicate stable or consistent organization within the data, while large fluctuations or rapid rises in topological features could indicate unstable, or very complex, representations. In this paper, persistence diagrams and Betti curves are used for analysing the embedding representations of Transformer models. These visualization techniques offer some important insights about the structural behaviour of learned representations and help to identify patterns which signal the occurrence of overfitting, during the training.

7. Benefits of Topological Analysis in Deep Learning

Using Persistent Homology to analyse Transformer models comes with many advantages. First, topological analysis can be used to determine the structural organization of embeddings, and give you more insight into the representations of the model itself.

Second, this is model agnostic, and therefore we can apply this method to different types of neural network architectures. Thus, this method is not only limited to Transformer models, but is also useful for analysing other deep learning models.

Third, and last but still a very important one, overfitting can be identified early on thanks to topological analysis. Structural instability may precede traditional indicators, such as training loss and validation loss, of a decrease in performance in the embedding space. Additionally, the Persistent Homology offers the potential for analyses at multiple scales, that is, to study connectivity at different levels as well as structural patterns in the data.

Due to all these features, Topological Data Analysis (TDA) shows to be an effective and promising tool to understand the structure and learning dynamics of the deep neural networks.

Conclusion:

This work shows the power of Persistent Homology from Topological Data Analysis (TDA) as a powerful, model indifferent procedure for analysing changing geometric structure of Transformer embeddings through training. By using Vietoris-Rips filtrations that we compute based on pairwise embedding distances, we traced the birth and the persistence of 0D (connected components) and 1D (loops) features using persistence diagrams and Betti curves. Our analysis shows a quite distinct topological signature of overfitting: the growth of short-lived and unstable features that point towards noise memorization compared to stable and long persisting structures in well generalised models.

These findings provide complementary information to traditional metrics such as the training-validation loss gap, where they give multi-scale interpretable insights into representation learning dynamics that allow early detection of overfitting - before performance degradation becomes apparent. The proposed TDA pipeline not only improves the interpretability and training stability of Transformer type architecture but also goes further, by extending to deep learning applications in general, it paved the way to topology-informed regularizations techniques and real-time monitoring techniques in large-scale NLP systems. Future work might investigate higher dimension features, might be its integration to attention maps, automated thresholds for topological alerts in production training pipelines etc.

References:

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762
2. Rathore, A., Zhou, Y., Srikumar, V., & Wang, B. (2023). TopoBERT: Exploring the topology of fine-tuned word representations. Information Visualization, 22(3), 186–208. https://doi.org/10.1177/14738716231168671
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
4. Fugacci, U., Scaramuccia, S., Iuricich, F., & De Floriani, L. (2016). Persistent homology: A step-by-step introduction for newcomers. STAR Track Short Papers, 2016. https://doi.org/10.2312/stag.20161358
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). NAACL-HLT.
7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
8. Kostenok, E., Cherniavskii, D., & Zaytsev, A. (2023). Uncertainty estimation of transformers' predictions via topological analysis of the attention matrices (arXiv preprint arXiv:2308.11295). https://arxiv.org/abs/2308.11295
9. Chazal, F., & Michel, B. (2021). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. Frontiers in Artificial Intelligence, 4, 667963. https://doi.org/10.3389/frai.2021.667963
10. Carlsson, G., & Vejdemo-Johansson, M. (2021). Topological data analysis with applications. Cambridge University Press.
11. Kostenok, E., Cherniavskii, D., & Zaytsev, A. (2023). Uncertainty estimation of transformers' predictions via topological analysis of the attention matrices (arXiv preprint arXiv:2308.11295). https://arxiv.org/abs/2308.11295
12. Wasserman, L. (2018). Topological data analysis. Annual Review of Statistics and Its Application, 5(1), 501–532. https://doi.org/10.1146/annurev-statistics-031017-100045
13. Fugacci, U., Scaramuccia, S., Iuricich, F., & De Floriani, L. (2016). Persistent homology: A step-by-step introduction for newcomers. STAR Track Short Papers, 2016. https://doi.org/10.2312/stag.20161358
14. Ghrist, R. (2008). Barcodes: The persistent topology of data. Bulletin of the American Mathematical Society, 45(1), 61–75. https://doi.org/10.1090/S0273-0979-07-01191-3
15. Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2003). Topological persistence and simplification. Discrete & Computational Geometry, 28(4), 511–533. https://doi.org/10.1007/s00454-002-2885-2
16. Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
18. Fefferman, C., Ivanov, S., Lassas, M., & Narayanan, H. (2023). Fitting a manifold to data in the presence of large noise (arXiv preprint arXiv:2312.10598). https://arxiv.org/abs/2312.10598
19. Edelsbrunner, H., & Harer, J. (2010). Computational topology: An introduction. American Mathematical Society. https://doi.org/10.1090/gsm/110
20. Uchendu, A., & Le, T. (2024). Unveiling topological structures in text: A comprehensive survey of topological data analysis applications in NLP (arXiv preprint arXiv:2411.10298). https://arxiv.org/abs/2411.10298
21. Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X
22. Fugacci, U., Scaramuccia, S., Iuricich, F., & De Floriani, L. (2016). Persistent homology: A step-by-step introduction for newcomers. STAR Track Short Papers, 2016. https://doi.org/10.2312/stag.20161358

23. Chazal, F., De Silva, V., Glisse, M., & Oudot, S. (2016). The structure and stability of persistence modules (SpringerBriefs in Mathematics, Vol. 10). Springer. https://doi.org/10.1007/978-3-642-33259-6