APCNCS2026: Asia-Pacific Conference on Networks and Complex Systems
10-12 June 2026, Nanyang Technological University, Singapore

Applying Transformer Models to Predict Topological Features of AI Citation Networks

Speaker: Yifei Li

Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

Keywords: transformers, Machine Learning, Betti numbers, topology, complex systems

Extended Abstract

Motivation.
In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have become the most strategic modern technology. It is being pursued by various countries in Southeast asia and Europe, along with the US and China. To have an advantage in AI and ML and maintain leverage in critical areas, we need to predict how future AI research trends evolve. A possible way to achieve this is to scrape AI bibliographic records and construct AI bibliographic coupling networks, characterize their topological and geometrical features, and predict the future evolution of these features. To do that, we need to resort to topological data analysis and tools developed from the Science-of-science field. One such tool is topological data analysis, which has attracted attention in the last decade. More recently, the geometry of data has also become very appealing [1], with applications found in community detection and geometric deep learning. A combination of these tools, with topology describing the constitution of the backbone of the data structure, and geometry adding additional information on top of that backbone, provides a more comprehensive picture of the structure of the data and its evolution.

In Science-of-science [2, 3], citation network is an useful and robust tool in elucidating how research topics evolve throughout the years. Prediction of the future evolution of these networks currently remains a major challenge. Approaching the problem from the perspective of a complex scientist, we believe persistent bibliographic coupling networks (BCN) will form first at different spatial scales, in which information processes across different spatial and temporal scales. These single information processing events can persist, or form cascades that emerge at other spatiotemporal scales.

In this work, we propose to train a transformer model by using the number of papers in each top BCNs and their corresponding Betti number β_n, (Figure 1) at weights $w = 5$, to see if the parameters and hyperparameters after training can give us an acceptable prediction accuracy. We then use the trained parameters to predict β_n of later years, or other weights such as $w = 4$ or 6 for the same time span, to conclude whether a cross-scale prediction is achievable.

Approach and Methodology.
The bibliographic records are first downloaded from the Web of Science website, using "artificial intelligence" and "neural networks" as keywords, which consist of nearly 1.2 million records from 1985 to 2023. We then compute BCNs for each year at different weights w, where BCNs can be identified by using a simple partitioning algorithm. For each BCN in each year, we compute its size N and β_n, $n = 0, 1, 2$. After which, the data is built into a feature vector that can be fed into a transformer model for training and prediction. Due to this distillation of

the BCNs down into its topological features, the amount of trainable data and the size of the training dataset is extremely small (n ≈ 30). Overfitting is thus a major concern as the number of available trainable parameters far exceeds that of the size of the dataset itself. A significant amount of effort is spent on solving the issue of overfitting and attempts to achieve generalization both on the same scale or across scale (multi scale prediction). To combat this, a small transformer model was built, to keep the number of trainable parameters low. This also has the added benefit of allowing quicker training and iterated testing to tune hyperparameters. A custom loss function is also written to allow better capturing of the desired features. The feed forward layer was also built with a different dimensionality instead of the proposed 4x by Vaswani et.al[4], in order to avoid overfitting on the extremely tiny dataset. The transformer model is then trained using the prepared bibliographic data up to 2018 as training sets, with the subsequent years of 2019 to 2023 as validation sets for to check for the prediction accuracy, before using the trained models to make predictions of the evolution of the size of top BCNs, and their β_ns in later years.

Results. We have computed the number of top 10 BCNs and their respective β_ns from 1991 to 2018 and used the data for fitting the transformer model's weights. It predicts the subsequent year's β_n based on input data. The transformer currently has a slight tendency to underestimate the rapidly exploding field when checked against the validation data from 2019 to 2023. The small data size means that a small number of training epochs (around 25) is sufficient before the model starts to overfit. We are currently working on improving the ability to capture and predict when the later years involves a field that is expanding rapidly expanding, starting with the specific $w = 5$ scale. Another test is to use the fitted model at $w = 5$ to predict nearby $w = 4$, and 6 BCNs results, before applying transfer learning to compute $w = 4$ and 6 results, and compare them with the BCNs features. We will then benchmark the performance and accuracy of applying transfer learning before giving the concluding remarks.

Conclusions and Outlook.

Besides predicting β_n at subsequent years, we are also looking into the issue of spatial stacking when we first trained and learned the features at $w = w'$, we suspect it will carry some characteristics over to $w = w' - 1$. While investigating the spatial stacking scenario, we will apply transfer learning to test its performance. If we find out that the parameters trained at $w = w'$predict wrong β_n results at $w = w' - 1$, we will then initiate the translation module (an encoder-decoder transformer model) and benchmark the performance against the β_n results predicted from using the trained parameters, and β_n results predicted from parameters from applying transfer learning.

(A)

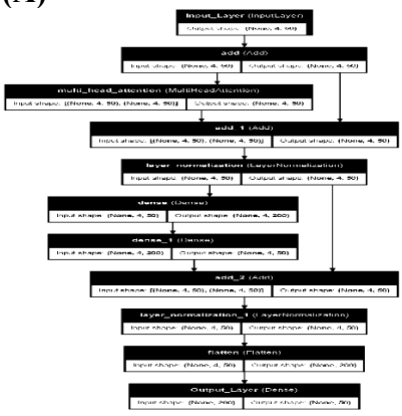

Figure 1. (A) An illustrative plot of the specific architecture generated from `pydot` python package. Specific modifications can also be captured by this package.

References

1. Yadav, Y. and Xia, K., *A roadmap for curvature-based geometric data analysis and learning.* arXiv preprint arXiv:2510.22599, 2025.
2. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., and Uzzi, B., *Science of science.* Science, **359**(6379): p. eaao0185, 2018.
3. Wang, D. and Barabási, A.-L., *The science of science*. Cambridge University Press, 2021.
4. Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30, 5998-6008.