Overview of Graph ViT/MLP-Mixer

Analysis of Complex Networks
University of Luxembourg

Anton Zaitsev
anton.zaitsev.001@student.uni.lu

Overview

Graph ViT/MLP-Mixer [1] is a novel graph neural network (GNN) architecture inspired by Vision Transformers (ViT) and MLP-Mixer models originally developed for computer vision tasks. The general process involves extracting patches from the input data, encoding these patches using a GNN-based patch encoder (e.g., Graph Convolutional Network (GCN), Graph Attention Network (GAT), or a Graph Transformer (GT)), then applying Mixer layers (e.g., MLP or gMHA) to the patch embeddings. Afterward, global average pooling is performed, and the resulting global embedding is passed through a fully connected layer for the final prediction.

Comparison with GraphSAGE, GAT, and GCN

GraphSAGE, GAT, and GCN are message-passing GNNs that rely on iterative aggregation and transformation of information from neighbors of a node. Message-passing GNNs compute node representation by aggregating the local 1-hop neighborhood information. Then, by using L layers, the model can gather information from nodes up to L hops away. This approach increases models ability to capture more complex and distant relationships and lets nodes that are farther apart share information, but suffers from over-smoothing. In contrast, Graph ViT/MLP-Mixer transforms the graph into patches and then uses Mixer layers to directly model global relationships, thus capturing long-range dependencies without solely relying on iterative neighbor aggregation.

Capabilities and Limitations

Capabilities:

- Can be used for any computer vision, natural language processing, and graph task by considering data as tokens. For images tokens are image patches; for NLP tokens are words or subwords; for graphs tokens are nodes (or node features). This creates a "unified architecture that can potentially benefit cross-over domain collaborations to design better networks" [1].
- Efficiently captures long-distance dependencies between nodes in a graph and "mitigates the issue of over-squashing" [1], addressing poor long-range dependency of MP-GNNs.
- Low computational cost, offering "better speed and memory efficiency with a complexity linear to the number of nodes and edges, surpassing the related Graph Transformer and expressive GNN models" [1].
- Has high expressive power: "... high expressivity in terms of graph isomorphism as they can distinguish at least 3-WL non-isomorphic graphs" [1].
- Addresses the issue of overfitting for ViT/MLP-Mixer architectures by augmenting the graph data. The general idea of the augmentation is to randomly drop a small set of edges before applying the partitioning algorithm, thus yielding slightly different graph partitions and patches each epoch.

Limitations:

- The effectiveness of the model depends on the structural complexity of the graph: "... achieving such a successful generalization is challenging given the irregular and variable nature of graphs" [1].
- Deciding how to first represent data as a graph and then to split a graph into meaningful and consistent patches is non-trivial and may affect model performance. Additionally, preserving local and global positional information in graphs is difficult, compared to some other modalities, such as image data.

Conclusion

Graph ViT/MLP-Mixer is a novel GNN architecture inspired by recent advances in computer vision. It effectively addresses several challenges assossiated with message-passing GNNs while maintaining computational and memory efficiency.

References

[1] Xiaoxin He, Bryan Hooi, Thomas Laurent, Adam Perold, Yann LeCun, and Xavier Bresson. A generalization of vit/mlp-mixer to graphs. In *International Conference on Machine Learning*, pages 12724–12745. PMLR, 2023.