Mathematical Formulation of DAG-ViT: A Vision Transformer with Adaptive Graph Reasoning

Veer Patel
veer.patel1801@gmail.com

***Abstract*—We present the mathematical formulation of DAG-ViT, a novel Vision Transformer framework that integrates adaptive token selection with graph-based reasoning for structured image analysis. Unlike conventional ViTs that use fixed-grid patch tokenization, DAG-ViT dynamically identifies informative image regions and models their relationships through a learnable Directed Acyclic Graph (DAG). We define the complete pipeline using formal notations, including adaptive token extraction, DAG-based graph learning, acyclicity constraints, graph convolution, and Transformer-based reasoning. This formulation offers a principled and interpretable mechanism to model spatial dependencies while improving computational efficiency. While we do not present empirical results in this work, we lay the theoretical foundation for future implementations in high-impact domains such as medical imaging and satellite-based analysis.**

***Index Terms*—Vision Transformer, Graph Neural Network, Directed Acyclic Graph, Adaptive Tokenization, Medical Imaging, Satellite Image Analysis**

I. Introduction

Vision Transformers (ViTs) have emerged as a transformative class of models in computer vision by leveraging self-attention to capture long-range dependencies in image data. Originally introduced in natural language processing, the transformer architecture was adapted to vision tasks by Dosovitskiy et al. [1], demonstrating competitive performance on large-scale image classification benchmarks. Despite their success, ViTs face significant challenges in terms of computational efficiency, interpretability, and adaptability—especially in applications requiring spatial precision and structured reasoning.

A core limitation of standard ViTs lies in their use of uniform patch tokenization, where the input image is split into fixed-size patches and processed equally regardless of visual importance. This approach is inefficient and often fails to emphasize semantically meaningful regions. Moreover, ViTs typically model global attention among all patches, overlooking the inherent spatial or semantic relationships between localized regions.

To address these challenges, recent advances have explored adaptive token selection strategies—such as TokenLearner and DynamicViT—to dynamically identify informative regions and reduce token redundancy. Parallelly, Graph Neural Networks (GNNs) have proven effective in modeling structural relationships between image components, enabling improved reasoning in tasks like scene understanding and medical diagnosis. However, existing GNN-based methods often depend on predefined graphs or heuristically constructed relationships, which limits their flexibility and generalization.

In this work, we present a formal mathematical formulation of DAG-ViT, a novel Vision Transformer architecture that unifies adaptive token selection with relational modeling through a learnable Directed Acyclic Graph (DAG). Unlike conventional ViTs, DAG-ViT dynamically selects a fixed number of informative tokens and models their dependencies using a directed, acyclic graph structure. This design offers both efficiency and interpretability by enabling localized focus and structured reasoning.

Our contribution is primarily theoretical: we define each component of DAG-ViT using rigorous mathematical expressions. This includes:

- Adaptive token extraction via soft attention masks over CNN feature maps.
- Graph construction through a learnable adjacency matrix with a differentiable acyclicity constraint.
- DAG-based message passing using graph convolutional updates.
- Transformer-based global reasoning over refined token embeddings.
- A total loss formulation combining classification and DAG regularization.

By formalizing DAG-ViT, we aim to provide a foundational blueprint for future implementations and empirical studies. Our framework is particularly suited for high-stakes domains—such as medical imaging and satellite analysis—where localized precision and relational structure are crucial.

This paper serves as a theoretical and mathematical groundwork for DAG-ViT, highlighting its potential to bridge the gap between adaptive attention and graph-based visual reasoning.

II. Related Work

The mathematical foundation of DAG-ViT is informed by advancements in Vision Transformers, adaptive tokenization strategies, graph neural networks, and acyclic graph learning. In this section, we review the theoretical works and architectural ideas that underpin our formulation, emphasizing conceptual influences rather than system-level performance.

A. Vision Transformers (ViTs)

ViTs represent a paradigm shift from convolutional models by using self-attention over image patches. Dosovitskiy et al. [1] formalized the ViT architecture by treating non-overlapping image patches as input tokens to a standard transformer. Follow-up works such as DeiT [2] and Swin

Transformer [3] contributed new variations for data efficiency and hierarchical modeling.

While these models inspired the attention-based formulation used in our framework, their reliance on fixed-grid tokenization motivates the development of a more adaptive token mechanism in DAG-ViT. Our formulation explicitly addresses this by introducing a learnable, content-aware token selection process.

B. Adaptive Token Selection

The idea of adaptive tokenization has emerged as a response to the inefficiencies in processing uniformly partitioned patches. TokenLearner [4] introduced a differentiable module to generate dynamic attention maps, while DynamicViT [5] implemented token pruning to reduce computational load during inference. Structure-Preserving Tokenization (SPT) [6] and Adaptive ViT (A-ViT) [7] also explored semantic and spatial saliency for token generation.

DAG-ViT builds on these works by formulating token selection as an attention-weighted integration over spatial features, enabling differentiable token learning guided by the image content. Unlike prior approaches, we additionally model relationships among selected tokens through a directed graph.

C. Graph Neural Networks in Vision

Graph-based modeling allows for explicit representation of relationships among image regions or objects. Foundational works such as Graph Convolutional Networks (GCN) [8] and Graph Attention Networks (GAT) [9] introduced formal methods for learning from graph-structured data, using spectral filtering and attention-based neighborhood aggregation, respectively.

In vision, GNNs have been used to model spatial relationships in videos [10] and object dependencies in scenes [11]. Our approach leverages the expressiveness of GNNs by defining a directed acyclic graph over adaptively selected tokens, allowing relational reasoning to be encoded directly into the transformer pipeline. Unlike static graph approaches, the DAG in our formulation is fully learnable and conditioned on the input image.

D. Acyclic Graph Constraints and DAG Learning

A central novelty in DAG-ViT is the use of a learnable adjacency matrix constrained to form a Directed Acyclic Graph (DAG). This draws from prior work on structure learning, particularly the NOTEARS method [12], which introduced a differentiable loss for enforcing acyclicity using trace-based matrix conditions. Extensions like DAGNN [13] and Contrastive DAG Learning [14] have further explored acyclicity in the context of causal inference and hierarchical classification.

We adapt the acyclicity constraint into our token graph formulation by mathematically enforcing that no cycles exist in the learned adjacency matrix. This provides a principled foundation for interpretable information flow across spatial tokens.

E. Theoretical Applications in Medical and Satellite Imaging

Although our paper focuses on theoretical formulation, the design of DAG-ViT is motivated by domains where interpretability, spatial reasoning, and structural modeling are critical—such as medical imaging and satellite analysis.

In medical imaging, transformer-based models have shown promise in capturing anatomical dependencies [15], and datasets such as CheXpert [16] and MIMIC-CXR [17] provide strong benchmarks for clinical visual tasks. Graph-based representations have also proven useful in modeling organ and lesion relationships.

In remote sensing, CNN and ViT-based systems like DeepSat [18] and TransUNet variants [19] have highlighted the need for spatial reasoning in environmental monitoring. Graph-based map extraction [20], [21] demonstrates the utility of structured modeling for high-resolution imagery.

By formulating DAG-ViT mathematically, we provide a flexible foundation that can be adapted to these domains in future implementations.

III. Methodology

We present the mathematical formulation of DAG-ViT, a framework that combines adaptive token selection and directed acyclic graph reasoning within a vision transformer architecture. The formulation aims to model spatial dependencies and relational structure in visual data, while preserving computational tractability and interpretability. The framework comprises the following components:

1) CNN-based feature extraction
2) Adaptive token generation using attention-weighted pooling
3) Learnable DAG-based graph construction with acyclicity constraints
4) Graph convolution for relational reasoning
5) Transformer-based encoding for global context modeling

Each component is defined below using formal notation.

A. Feature Extraction

Let $I \in \mathbb{R}^{H \times W \times C}$ denote an input image. A convolutional backbone $\text{CNN}(\cdot)$ maps this input to a spatial feature map:

$$F = \text{CNN}(I), \quad F \in \mathbb{R}^{H' \times W' \times d}$$

Here, H' and W' are the spatial dimensions, and d is the feature depth.

B. Adaptive Token Extraction

We define K adaptive tokens $\{T_1, \ldots, T_K\}$ using learnable soft attention masks over the feature map F. For each token $T_i \in \mathbb{R}^d$:

$$M_i(p) = \frac{\exp(f_{\text{att}}(F(p))_i)}{\sum_{p'} \exp(f_{\text{att}}(F(p'))_i)} \quad \text{(attention mask)}$$

$$T_i = \sum_p M_i(p) \cdot F(p) \quad \text{(adaptive token)}$$

where f_{att} is a learnable MLP. This formulation generalizes fixed patch extraction by weighting pixel contributions dynamically.

C. DAG Construction and Graph Learning

We construct a directed acyclic graph over the K tokens by learning a weighted adjacency matrix $A \in \mathbb{R}^{K \times K}$, where:

$$A_{ij} = \begin{cases} \frac{\exp(\text{MLP}([T_i;T_j]))}{\sum_{k>i} \exp(\text{MLP}([T_i;T_k]))}, & \text{if } j > i \\ 0, & \text{otherwise} \end{cases}$$

This ensures a strict partial order: edges are allowed only from lower- to higher-indexed tokens, enforcing a topological ordering.

D. Acyclicity Constraint

To regularize the graph structure and ensure it is acyclic, we adopt a trace-based penalty inspired by the NOTEARS method:

$$\mathcal{L}_{\text{DAG}} = \sum_{k=2}^{K} \text{tr}(A^k)$$

This formulation penalizes cycles of length ≥ 2, and in practice, powers up to $k = 3$ are typically sufficient.

E. Graph Convolution and Message Passing

The refined tokens $\hat{T}_i$ are computed using directed message passing over the DAG:

$$\hat{T}_i = \sigma\left(\sum_{j \in \mathcal{N}(i)} A_{ij} W T_j + b_i\right)$$

Here, $\mathcal{N}(i)$ is the set of predecessors of node i, $W \in \mathbb{R}^{d \times d}$ is a shared learnable weight matrix, and σ is a non-linear activation function (e.g., GELU).

F. Transformer Encoding

The updated tokens $\{\hat{T}_1, \ldots, \hat{T}_K\}$ are passed through a multi-head self-attention encoder:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$

This models global dependencies among the relationally refined tokens.

G. Output Representation

The final representation is obtained by average pooling:

$$\tilde{T} = \frac{1}{K}\sum_{i=1}^{K} \hat{T}_i$$

which can be passed to a downstream classifier or used as a latent embedding, depending on the task.

H. Training Objective

The total loss function combines the task-specific loss (e.g., classification cross-entropy) with the DAG regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{DAG}}$$

where λ is a scalar hyperparameter controlling the importance of enforcing acyclicity.

I. Implementation Considerations (Optional)

While this paper is theoretical, all components are formulated to be differentiable and implementable in standard deep learning frameworks (e.g., PyTorch). Operations such as `einsum`, matrix exponentiation, and masked softmax are sufficient to implement the model end-to-end.

Notation Clarification

To enhance clarity and reproducibility, we define the key notations used in this paper:

- $I \in \mathbb{R}^{H \times W \times C}$: Input image with height H, width W, and channels C.
- $F \in \mathbb{R}^{H' \times W' \times d}$: Feature map extracted from CNN backbone with spatial dimensions H', W' and feature depth d.
- $T_i \in \mathbb{R}^d$: The i^{th} adaptive token vector.
- $M_i(p)$: Soft attention mask for token T_i at position p.
- f_{att}: A lightweight MLP (multi-layer perceptron) that computes spatial attention weights; referred to as the "feature attention MLP".
- $[T_i; T_j]$: Concatenation of token vectors T_i and T_j, used as input to the MLP when computing edge weights in the adjacency matrix.
- A_{ij}: Entry in the learnable adjacency matrix representing the directed edge weight from T_j to T_i.
- $\hat{T}_i$: Refined token after DAG-based graph convolution.
- $\mathcal{N}(i)$: Set of nodes with incoming edges to node i.
- $\sigma(\cdot)$: Non-linear activation function, such as GELU or ReLU.
- $\mathcal{L}_{\text{DAG}}$: Differentiable loss function enforcing acyclicity on the graph.
- $\mathcal{L}_{\text{total}}$: Final training objective combining classification loss and DAG regularization.

IV. Applications

The mathematical formulation of DAG-ViT suggests broad applicability to vision tasks that require both localized attention and structured relational reasoning. In particular, its design supports adaptive region selection and graph-based dependency modeling, making it theoretically suitable for domains such as medical image analysis and satellite remote sensing.

A. Medical Image Analysis

Medical imaging is characterized by challenges such as high spatial resolution, class imbalance, and the presence of clinically relevant features that are sparse and spatially interdependent. DAG-ViT's formulation enables the selection of semantically meaningful regions and models their interactions through a directed graph structure.

Potential Advantages:

- Enables attention over anatomically relevant regions (e.g., lesions, joints, organ boundaries).
- Facilitates structured reasoning across regions exhibiting correlated pathology (e.g., fracture propagation).
- Improves theoretical interpretability by explicitly modeling dependencies through directed edges.
- Highlights subtle anomalies by prioritizing high-utility areas in token selection.

Theoretical Use Cases:

- **Fracture analysis in radiographs:** Identifying discontinuities and relational deformation across bones.
- **Tumor boundary modeling in CT/MRI:** Capturing spatial and structural associations between masses and surrounding tissues.
- **Organ-wise relational segmentation:** Learning topological relationships between adjacent organs or compartments.
- **Retinal anomaly detection:** Reasoning over fine-grained structures such as blood vessels and lesions in fundus images.

B. Satellite Image Analysis

Satellite imagery poses challenges including heterogeneous scales, sparsity of targets, and topological dependencies between geographical entities. DAG-ViT's formulation allows efficient processing of high-resolution scenes by adaptively sampling critical regions and modeling their spatial dependencies via directed graphs.

Potential Advantages:

- Enables dynamic region selection in scenes with large background-to-foreground imbalance.
- Graph reasoning captures spatial and functional relationships (e.g., roads linking buildings, forest edge transitions).
- Reduces computational overhead by avoiding uniform patch processing.
- Provides structural interpretability through explicit graph topology over regions of interest.

Theoretical Use Cases:

- **Urban structure modeling:** Inferring relationships between expanding urban zones.
- **Forest degradation monitoring:** Identifying and correlating patterns of vegetation loss across time-series frames.
- **Hydrological feature tracking:** Capturing seasonal or climate-induced changes in lakes and rivers.
- **Post-disaster impact mapping:** Reasoning over connectivity disruptions (e.g., broken bridges, flooded roads).

While empirical evaluation is outside the scope of this formulation paper, the structured design of DAG-ViT provides a strong theoretical foundation for future exploration in these high-impact areas. Its mathematical properties—adaptive focus and acyclic graph structure—may be especially useful for tasks requiring explainability, spatial prioritization, and topological reasoning.

V. Conclusion

In this paper, we introduced a detailed mathematical formulation of **DAG-ViT**, a novel framework that integrates adaptive token extraction with directed acyclic graph-based reasoning for vision tasks. The formulation defines how dynamically selected visual tokens can be connected through a learnable DAG structure to enable structured information flow, and how such a graph can be enforced using a differentiable acyclicity constraint.

Our contribution is grounded in a modular design that combines well-established techniques from CNN feature extraction, soft attention mechanisms, acyclic graph learning, and Transformer encoding. By rigorously formalizing the forward pass, adjacency matrix construction, DAG loss, and training objective, we provide a foundation for interpretable, sparse, and efficient vision models.

While this work is theoretical, it lays the groundwork for future implementation and experimental validation of DAG-ViT across real-world datasets. The proposed framework is especially promising for applications where spatial prioritization and topological reasoning are essential, such as medical image analysis and satellite scene understanding.

We hope this formulation will inspire future research at the intersection of graph theory and vision transformers, enabling further exploration into structured visual understanding with strong theoretical underpinnings.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021, pp. 10 347–10 357.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[4] M. S. Ryoo, A. P. Kim, J. Xia, D. So, and Q. V. Le, "Tokenlearner: What can 8 learned tokens do for images and videos?" *arXiv preprint arXiv:2106.11297*, 2021.

[5] Y. Rao, W. Zhao, Z. Tang, J. Zhou, C.-J. Hsieh, and J. Lu, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," in *Advances in Neural Information Processing Systems*, 2021, pp. 23 745–23 758.

[6] Y. Li, Z. Tang, S. Zhao, L. Lin, H. Peng, X. Xie, Q. Wu, and T. Huang, "Revisiting tokenization for vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 790–17 799.

[7] M. Muller, A. Dosovitskiy, and T. Brox, "Adaptive token sampling for efficient vision transformers," *arXiv preprint arXiv:2106.02852*, 2021.

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[9] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[10] X. Wang, R. Girshick, A. Gupta, and K. He, "Videos as space-time region graphs," in *European conference on computer vision*, 2018, pp. 399–417.

[11] J. Qi, Y. Yang, Y.-Z. Song, and T. Xiang, "Learning long-term dependencies for action recognition with a biologically-inspired deep network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6626–6635.

[12] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[13] W. Yu, W. Cheng, Z. Yu, Y. Song, and F. Nie, "Dagnn: Deep attentive graph neural network," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019.

[14] W. Ke, X. Zheng, J. Li, Z. Huang, H. Zhao, and J. Liu, "Learnable dags via contrastive regularization," *arXiv preprint arXiv:2206.07662*, 2022.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[16] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, C. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.

[17] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr: A large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[18] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "Deepsat: A learning framework for satellite imagery," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015, pp. 1–10.

[19] A. Doshi, R. Bhatt, P. Deshmukh, and A. Shah, "Transformers in remote sensing: A survey," *arXiv preprint arXiv:2202.01211*, 2022.

[20] J. Li, Y. He, H. Zhao, Z. Zhang, X. Huang, and J. Kautz, "Topological map extraction from overhead images," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 153–170.

[21] H. Zhu, Y. Zhang, H. Zhang, J. Wu, and J. Xiao, "Map alignment with graph neural networks," *arXiv preprint arXiv:2004.10638*, 2020.