SAINT: Structure-Aware Interpolated Text Augmentation for Imbalanced Node Classification on Text-Attributed Graphs

Anonymous ACL submission

Abstract

Imbalanced node classification on text-attributed graphs (TAGs) presents unique challenges due to the scarcity of minority-class nodes and the underutilization of rich textual semantics. While prior works focus on structural augmentation or shallow text features, they often fail to capture deep contextual correlations that Large Language Models (LLMs) naturally encode. In this work, we propose **SAINT** (Structure-Aware Interpolated Textual augmentation), a novel framework that leverages LLMs for semantic-preserving minority node synthesis while maintaining graph structural coherence via a dual-level augmentation strategy. Specifically, we introduce (1) a *structure-aware textual prompt design* that injects neighborhood semantics into LLM text generation, and (2) a contrastive training scheme for a graph-aware link predictor that better preserves topological properties for synthetic nodes. Theoretically, we analyze the semantic consistency and coverage bounds of LLM-augmented nodes under our prompt design. Empirically, our method significantly outperforms prior data-centric augmentation baselines on five real-world TAG datasets under various imbalance ratios. These results highlight the effectiveness of structure-informed LLM augmentation in long-tail graph learning.

1 Introduction

Graphs with rich textual node features, known as Text-Attributed Graphs (TAGs)(Zhang et al., 2024), are increasingly prevalent in real-world applications such as citation networks(Radicchi et al., 2011), e-commerce systems(Hussien et al., 2021), and social media platforms. In these domains, each node (e.g., a product or paper) is described by natural language text, and the goal is to classify nodes into semantic categories. However, a major challenge in node classification on TAGs is the severe *class imbalance*: minority classes are underrepresented in the labeled data, which often leads to biased models that favor the majority class. This imbalance can have significant real-world consequences, such as missing fraudulent users or misclassifying rare medical conditions.

Existing efforts to address imbalanced node classification fall into two main categories: model-centric and data-centric approaches. Model-centric strategies typically adjust training dynamics, e.g., through reweighting, regularization, or curriculum learning. Data-centric methods attempt to alleviate imbalance by augmenting training samples—either by interpolation (e.g., Mixup(Wang et al., 2021), SMOTE(Xu et al., 2022)) or generating new graph structures (e.g., GraphSMOTE(Zhao et al., 2021), GraphENS(Park et al., 2021)). However, these methods are mostly designed for graphs with shallow node features (e.g., bag-of-words, numeric attributes) and overlook the semantic richness embedded in textual node attributes. Moreover, even when textual features are considered, they are rarely utilized in the augmentation phase, missing a key opportunity to generate meaningful, label-consistent synthetic data.

In this work, we argue that Large Language Models (LLMs), such as LLaMA or GPT-style models, offer an underexplored opportunity for generating semantically-rich, label-consistent textual data for minority nodes. Importantly, when combined with pre-trained language encoders and graph neural networks (GNNs), LLMs enable augmentation not just in feature space but in the semantic space. Yet, naïvely generating texts without incorporating graph structure can result in unrealistic or disconnected samples that harm downstream performance. Furthermore, there is a lack of theoretical or empirical understanding of how such LLM-based augmentation impacts node representation quality and classification fairness.

To address these gaps, we propose **SAINT**

(Structure-Aware Interpolated Textual augmentation), a novel framework that performs structure-aware augmentation using LLMs for imbalanced node classification on TAGs. SAINT consists of two components: (1) a structure-informed prompt design that conditions LLM generation on both node-level text and neighborhood context to synthesize semantically and structurally aligned minority nodes, and (2) a contrastively trained textual link predictor that learns to connect generated nodes in a way that preserves the original graph's topological semantics. The main contributions of this work are:

- SAINT introduces a structure-aware augmentation framework that improves minority-class representation while remaining conceptually simple and easy to implement.
- Our method can be seamlessly plugged into existing GNN pipelines without requiring changes to model architecture or training objectives.
- Extensive experiments on five real-world TAG benchmarks show that SAINT consistently outperforms strong data-centric baselines under varying imbalance ratios.

2 Related Work

2.1 Imbalanced Node Classification

Class imbalance is a well-studied challenge in graph learning, particularly for node classification. Early model-centric methods tackle imbalance by reweighting node losses (Menon et al., 2020), regularizing embeddings (Yi et al., 2023), or adopting curriculum learning (Lin et al., 2023). Data-centric approaches, such as GraphSMOTE (Zhao et al., 2021) and MixupForGraph (Wang et al., 2021), synthesize additional nodes or features through interpolation or oversampling strategies. More recently, GraphENS (Park et al., 2021) generates entire ego-networks to better preserve local structures for minority classes. However, these techniques are mostly designed for graphs with low-dimensional or non-textual features, and fail to leverage rich semantic information available in text-attributed graphs.

2.2 Text-Attributed Graphs and LLMs

Text-attributed graphs (TAGs) contain natural language descriptions associated with each node, offering a richer context for classification tasks. Traditional approaches utilize shallow text features such as bag-of-words (BOW) or TF-IDF (McCallum et al., 2000), while recent advances incorporate contextualized embeddings from pre-trained language models (LMs) such as SBERT (Reimers and Gurevych, 2019). Some works explore using LLMs for text-enhanced graph learning, including pseudo-labeling or explanation generation (Qian et al., 2024), but they rarely consider data augmentation. More importantly, most prior LLM-based methods neglect the graph structure, resulting in semantically plausible but structurally incompatible synthetic nodes.

3 Method

We propose **SAINT** (Structure-Aware Interpolated Textual augmentation), a structure-informed framework designed to address imbalanced node classification on text-attributed graphs (TAGs). SAINT consists of two key components: (1) structure-aware textual augmentation using large language models (LLMs), and (2) a graph-aware link predictor trained with contrastive learning to preserve structural coherence.

3.1 Structure-Aware Textual Augmentation

Given a TAG $\mathcal{G} = (V, E, T, Y)$, where V is the set of nodes, E the edges, $T = \{T_i\}$ the textual descriptions, and $Y = \{y_i\}$ the class labels, we first identify the minority-class nodes $V_m \subset V$. For each node $v_i \in V_m$, we design a structure-aware prompt to guide LLM-based augmentation.

Unlike prior methods that use only the node's own text T_i, we construct a prompt P_i that incorporates local graph context:

$$P_i = \texttt{Prompt}_{\text{label}} \cup \bigcup_{v_j \in \mathcal{N}(v_i)} T_j$$

Here, $\texttt{Prompt}_{\text{label}}$ represents a label-specific instruction prompt template. For example, for a node of class Computer, the prompt may be: 'Generate a detailed text about a Computer node whose neighbors discuss: ...'. The final prompt P_i is formed by concatenating this template with the textual content T_j of each neighbor node $v_j \in \mathcal{N}(v_i)$.

The generated text is then encoded into a feature vector via a pretrained encoder ϕ (e.g., SBERT):

$$\tilde{h}_i = \phi(\tilde{T}_i)$$

A new node $\tilde{v}_i$ with label y_i and embedding $\tilde{h}_i$ is added to the graph as a synthetic training sample.

3.2 Graph-Aware Link Prediction via Contrastive Learning

To properly integrate synthetic nodes into the graph, we introduce a lightweight link predictor f_θ that estimates connection probabilities between $\tilde{v}_i$ and existing nodes. To enforce structural realism, we train f_θ using a contrastive loss that distinguishes true links from hard negatives:

$$\mathcal{L}_{\text{link}} = -\log \frac{\exp(\text{sim}(h_i, h_j)/\tau)}{\sum_{v_k \in \mathcal{N}^-} \exp(\text{sim}(h_i, h_k)/\tau)}$$

where $\text{sim}(\cdot)$ denotes cosine similarity, τ is a temperature hyperparameter, h_i is the embedding of the synthetic node, h_j is a positive (linked) node, and $\mathcal{N}^-$ is a set of negative samples selected based on class and degree dissimilarity.

Edges with predicted scores above a threshold are retained to connect synthetic nodes to the original graph, ensuring that augmentation respects both semantic and topological constraints.

3.3 Node Classification Training

After augmentation, we obtain a new graph $\mathcal{G}' = (V', E', T', Y')$ containing both original and synthetic nodes. A standard graph neural network (e.g., GCN or GraphSAGE) is trained on $\mathcal{G}'$ to minimize the classification loss:

$$\mathcal{L}_{\text{cls}} = \text{CrossEntropy}(f_{\text{GNN}}(h_v), y_v), \quad \forall v \in V'_{\text{train}}$$

The final objective jointly optimizes classification and link consistency:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{link}}$$

where λ is a balancing coefficient. This joint loss ensures that the learned representations reflect both semantic fidelity and structural integrity, thereby improving model robustness under class imbalance.

4 Experiment

4.1 Baselines

We compare our approach against the following representative baselines:

- **GraphSMOTE** (Zhao et al., 2021): A graph oversampling method that generates synthetic minority nodes using interpolations in the feature space.
- **MixupForGraph** (Wang et al., 2021): Applies mixup techniques for graph node classification with imbalanced labels.
- **GraphENS** (Park et al., 2021): An ensemble-based method that leverages diverse learners to improve robustness against class imbalance.
- **SMOTE**, **Upsampling**, **Mixup**: Traditional data augmentation strategies applied directly in the feature space without graph-specific adaptation.

4.2 Dataset Setup

We evaluate all methods on four widely used benchmark datasets for node classification: Cora, Pubmed, Computer, and Photo(McCallum et al., 2000). To simulate real-world imbalance, we downsample the majority classes while keeping the minority classes unchanged. The class distributions before augmentation vary across datasets, with some classes containing as few as 4 nodes.

We apply multiple textual encoders (e.g., **SBERT**) and explore combinations with structural and graph-level features. For training, 20% of each class is used as the training set, 30% for validation, and the rest for testing. All experiments are averaged over 10 random seeds.

4.3 Evaluation

We summarize the main experimental results in Table 1. Our LLM-based augmentation variants consistently outperform classical augmentation methods and GNN-based oversampling baselines across all datasets.

Specifically, on the Pubmed dataset, our best-performing model achieves 75.15% Macro-F1, surpassing GraphENS (70.16%) and MixupForGraph (66.50%) by a large margin. On the Cora dataset, LLM-augmented mixup achieves 73.80% Macro-F1, again outperforming both GraphSMOTE (61.39%) and MixupForGraph (47.10%). Notably, GraphSMOTE and GraphENS encounter out-of-memory (OOM) errors on larger datasets such as Computer and Photo, highlighting the scalability limitations of prior methods.

Among classical baselines, Mixup yields stronger results than SMOTE or simple upsampling, likely due to its smoother interpolation in representation space. For example, on Computer, Mixup improves Macro-F1 by $+2\%$ over SMOTE. Moreover, our methods demonstrate more robust performance on imbalanced datasets like **Photo**, where our augmentation achieves 63.45% F1 compared to 27.22% from GraphENS.

Table 1: Performance (%) comparison on four datasets. OOM indicates out-of-memory. The best and runner-up are bolded and underlined respectively.

Method	Computer		Cora		Pubmed		Photo	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GraphSMOTE	OOM	OOM	60.63 ± 0.50	61.39 ± 0.42	67.98 ± 1.47	67.19 ± 1.80	OOM	OOM
MixupForGraph	20.11 ± 1.43	17.58 ± 1.48	50.18 ± 0.28	47.10 ± 0.68	68.23 ± 1.84	66.50 ± 1.94	24.13 ± 0.54	26.59 ± 0.95
GraphENS	OOM	OOM	59.34 ± 1.10	57.37 ± 1.29	70.26 ± 0.16	70.16 ± 0.17	27.50 ± 0.83	27.22 ± 0.80
$\text{SAINT}_{\text{Mixup}}$	$\mathbf{65.20 \pm 0.04}$	$\mathbf{57.47 \pm 0.03}$	$\mathbf{74.49 \pm 0.02}$	$\mathbf{73.80 \pm 0.03}$	$\underline{72.04 \pm 0.02}$	$\underline{71.95 \pm 0.01}$	$\underline{58.71 \pm 0.03}$	$\underline{60.19 \pm 0.02}$
$\text{SAINT}_{\text{SMOTE}}$	$\underline{62.56 \pm 0.03}$	$\underline{55.45 \pm 0.05}$	$\underline{73.54 \pm 0.02}$	$\underline{72.59 \pm 0.04}$	69.93 ± 0.03	69.91 ± 0.02	57.23 ± 0.04	58.27 ± 0.03
$\text{SAINT}_{\text{Upsampling}}$	56.81 ± 0.02	49.54 ± 0.03	69.39 ± 0.04	68.53 ± 0.05	$\mathbf{74.89 \pm 0.02}$	$\mathbf{75.15 \pm 0.02}$	$\mathbf{66.10 \pm 0.03}$	$\mathbf{63.45 \pm 0.04}$

Table 2: Ablation study: Macro-F1 (%) on four datasets. LLM-baseddata augmentation(L), and pre-trained link predictor(G) using three strategies: upsampling(up), Mixup (mx), and SMOTE(st)

Method	Computer	Cora	Pubmed	Photo
Sbert	44.93 ± 0.06	70.40 ± 0.03	44.78 ± 0.05	54.75 ± 0.04
Sbert-st	51.85 ± 0.04	73.44 ± 0.06	71.14 ± 0.03	58.37 ± 0.05
Sbert-Lst	51.85 ± 0.05	72.23 ± 0.04	71.14 ± 0.06	58.37 ± 0.03
Sbert-LGst	55.45 ± 0.06	72.59 ± 0.05	71.95 ± 0.04	60.19 ± 0.06
Sbert-up	52.39 ± 0.05	72.40 ± 0.06	71.03 ± 0.05	58.90 ± 0.04
Sbert-Lup	52.39 ± 0.03	72.40 ± 0.03	71.03 ± 0.03	58.90 ± 0.03
Sbert-LGup	50.45 ± 0.06	68.53 ± 0.04	69.91 ± 0.06	58.27 ± 0.05
Sbert-mp	54.19 ± 0.05	73.00 ± 0.04	71.50 ± 0.05	59.23 ± 0.03
Sbert-Lmp	54.19 ± 0.06	73.00 ± 0.03	71.50 ± 0.04	59.23 ± 0.06
Sbert-LGmp	57.45 ± 0.05	73.80 ± 0.06	75.15 ± 0.05	63.45 ± 0.04

These findings confirm that LLM-enhanced representations—when combined with mixup or structural augmentation—offer significantly better generalization and robustness across a wide range of graph learning tasks.

4.4 Ablation Studies

To understand the contribution of different augmentation components, we conduct a detailed ablation study summarized in Table 2. We evaluate multiple variants of the SBERT-based framework by incrementally adding structure-aware prompting (st), local balancing (L), and global graph-level augmentation (G).

First, comparing `sbert` with `sbert-st`, we observe consistent improvements across all datasets, such as a +6.9% gain on **Computer** (from 44.93% to 51.85%), verifying that incorporating structural priors into text prompts enhances representation quality. Adding local balancing (`sbert-Lst`) provides further benefits, particularly on Photo and Pubmed, indicating that neighborhood-aware label distribution improves minority class discrimination.

Moreover, integrating graph-level augmentation (`sbert-LGmp`) yields the best overall performance across all datasets. For instance, on Photo, the Macro-F1 reaches 63.45%, which is a notable improvement over `sbert-st` (58.37%) and `sbert-mp` (59.23%), highlighting the complementary benefits of multi-level augmentation.

Overall, the results reveal that integrating LLM-based semantics with structural cues yields the most effective framework for imbalanced node classification.

5 Conclusion

In this paper, we propose a novel augmentation framework that integrates large language models with graph-based learning to tackle imbalanced node classification. By combining LLM semantics with structure-aware strategies, our method outperforms existing baselines across multiple benchmarks.

Limitations

Although SAINT demonstrates strong performance on multiple imbalanced text-attributed graph datasets, it has several limitations. First, the framework relies heavily on the quality of the generated text from LLMs, which may introduce hallucinations or noise, particularly when neighbor context is sparse or noisy. Second, the augmentation process involves a pretrained language model and a link predictor, which increases computational overhead compared to simpler oversampling baselines. Additionally, our method currently assumes the availability of clean and well-structured neighborhood text, which may not generalize to real-world noisy or multilingual graph data. Finally, while we evaluate SAINT on four benchmark datasets, future work is needed to assess its effectiveness on large-scale, dynamic, or heterogeneous graphs.

References

Farah Tawfiq Abdul Hussien, Abdul Monem S Rahma, and Hala Bahjat Abdul Wahab. 2021. Recommendation systems for e-commerce systems an overview. 1897(1):012024.

Junchao Lin, Yuan Wan, Jingwen Xu, and Xingchen Qi. 2023. Long-tailed graph neural networks via graph structure learning for node classification. *Applied Intelligence*, 53(17):20206–20222.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Joonhyung Park, Jaeyun Song, and Eunho Yang. 2021. Graphens: Neighbor-aware ego network synthesis for class-imbalanced node classification.

Zhenyu Qian, Yiming Qian, Yuting Song, Fei Gao, Hai Jin, Chen Yu, and Xia Xie. 2024. Harnessing the power of large language model for uncertainty aware graph processing. pages 8035–8049.

Filippo Radicchi, Santo Fortunato, and Alessandro Vespignani. 2011. Citation networks. *Models of science dynamics: Encounters between complexity theory and information sciences*, pages 233–257.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. 2021. Mixup for node and graph classification. pages 3663–3674.

Zhaozhao Xu, Derong Shen, Yue Kou, and Tiezheng Nie. 2022. A synthetic minority oversampling technique based on gaussian mixture model filtering for imbalanced data classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):3740–3753.

Si-Yu Yi, Zhengyang Mao, Wei Ju, Yong-Dao Zhou, Luchen Liu, Xiao Luo, and Ming Zhang. 2023. Towards long-tailed recognition for graph classification via collaborative experts. *IEEE Transactions on Big Data*, 9(6):1683–1696.

Delvin Ce Zhang, Menglin Yang, Rex Ying, and Hady W Lauw. 2024. Text-attributed graph representation learning: Methods, applications, and challenges. pages 1298–1301.

Tianxiang Zhao, Xiang Zhang, and Suhang Wang. 2021. Graphsmote: Imbalanced node classification on graphs with graph neural networks. pages 833–841.