International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 13 Issue: 1
***Article Received:** 25 April 2025 **Revised:** 12 May 2025**Accepted:** 25 June 2025*

Self-Supervised Hierarchical Representation Learning for Multi-Dimension Context

Manali Jahagirdar
Departpment of Computer Engineering
SCTR'S Pune Institute of Computer Technology
Pune, India
msjahagirdar@pict.edu

Mukta Takalikar
Departpment of Computer Engineering
SCTR'S Pune Institute of Computer Technology
Pune, India
mstakalikar@pict.edu

Abstract—Self-supervised hierarchical representation learning offers an effective approach to capturing multi-dimensional context from unlabeled data. A key challenge in representation learning is integrating information from diverse aspects of the input, particularly when labeled data is limited. To address this, a novel strategy can be introduced that learns representations hierarchically, enabling the capture of context at varying levels of abstraction and across multiple dimensions. The process begins by modeling different contextual facets through component-specific representations, each capturing distinct semantic and structural attributes. A dynamic aggregation mechanism then combines these representations in a hierarchical manner, allowing information to propagate across levels of contextual abstraction. This enables the encoding of both fine-grained nuances and broader contextual dependencies. By leveraging self-supervised learning, the approach optimizes for inherent relationships within the multi-dimensional context, enabling the acquisition of robust representations from unlabeled data. This makes it particularly suitable for domains where labeled data is scarce or costly to obtain. Experimental results highlight the ability to learn rich, hierarchical representations that enhance performance on downstream tasks requiring deep contextual understanding. Key technical contributions include: (1) a context-aware masking strategy using Text Encoder for semantic recovery of masked fields, (2) a Hierarchical Model that fuses fine-grained tabular features with coarse-grained concepts and (3) a multi-stage training code base combining contrastive loss for cross-document alignment (RFP-bid pairs) and silhouette scores (from scikit-learn) to validate cluster coherence.

Keywords- Self-supervised learning; Hierarchical representation learning; Contrastive learning; Multimodal; Text classification; Document embedding; Tabular Data Processing

I. INTRODUCTION

Self-supervised learning (SSL) has become a crucial approach because annotating large datasets for machine learning can be too expensive or infeasible due to the sheer volume of data. The increasing volume and complexity of data necessitate effective strategies for extracting meaningful representations that capture the intricate relationships present within. A significant challenge lies in learning such representations without relying on extensive manual annotations, which can be a limiting factor in many real-world applications where data labeling is expensive or impractical due to scale. To overcome this limitation, self-supervised learning (SSL) has emerged as a powerful paradigm, enabling models to learn from unlabeled data by defining pretext tasks that utilize the inherent structure of the data itself to generate supervisory signals. By learning to solve these self-generated tasks, models can develop representations that encode valuable semantic information transferable to various downstream tasks.

Complementary to the advances in SSL is the recognition that many types of data exhibit inherent hierarchical structures across multiple dimensions. For instance, a document can be viewed as a hierarchy of sections, paragraphs, and sentences, each providing a different level of contextual information. Similarly, in visual scenes, objects are composed of parts, and scenes are composed of objects, forming a visual hierarchy. Hierarchical representation learning (HRL) aims to capture these multi-level structures, allowing models to reason at varying granularities of abstraction and to better understand the complex relationships within the data. By modeling these hierarchical relationships, the learned representations can become more robust and informative, leading to improved performance in tasks requiring a deep understanding of context.

The integration of self-supervised learning with hierarchical representation learning offers a promising avenue for addressing the challenges of learning from complex, multi-dimensional data where labeled examples are scarce.

By leveraging self-generated supervision to learn representations at different levels of abstraction, it becomes possible to develop more efficient and scalable learning frameworks with reduced dependence on human annotation. This combination is particularly relevant for understanding data where context is not monolithic but rather a composite of various interacting facets organized in a hierarchy.

One approach to realizing this synergy involves modeling different contextual aspects through representations that are specific to those components. These component-specific representations can be designed to capture distinct semantic and structural attributes inherent in the data. Subsequently, a dynamic aggregation mechanism can be employed to combine these representations in a hierarchical manner. This hierarchical aggregation allows information to flow and be integrated across different levels of contextual abstraction, enabling the encoding of both fine-grained details and broader contextual dependencies within the learned representations.

By grounding this learning process in self-supervision, the model can be trained to optimize for the intrinsic relationships that exist within the multi-dimensional context of the unlabeled data. This approach is particularly advantageous in domains where obtaining labeled data for every possible contextual nuance is impractical. The ability to learn rich, hierarchical representations in a self-supervised manner holds the potential to significantly enhance performance on downstream tasks that demand a comprehensive understanding of context across multiple dimensions.

Key considerations in developing such a framework include strategies for identifying and representing the different facets of context, designing effective hierarchical aggregation mechanisms, and formulating self-supervised objectives that drive the learning of meaningful and context-aware representations. Furthermore, validating the coherence and quality of the learned hierarchical representations is crucial for ensuring their utility in downstream applications. Addressing these aspects can lead to significant advancements in leveraging the wealth of unlabeled data for learning deep and contextually rich representations.

Hierarchical representation learning (HRL) is also gaining attention due to the inherent hierarchical structures present in many real-world data sources. For example, documents have a hierarchy of sections, paragraphs, and sentences. HRL aims to learn representations that capture these multi-level structures, allowing models to reason at different levels of granularity and potentially leading to more robust and informative representations.

Hyperbolic embedding is a technology that has been successfully applied to computer vision and natural language processing tasks because of the unique property of hyperbolic space: its exponential volume growth relative to the radius, unlike the polynomial growth in Euclidean space. Hyperbolic embedding has shown promise in modeling relational data, such as graphs, which often exhibit complex hierarchical structures. Various hyperbolic graph embedding models leverage the distinctive properties of hyperbolic space to model these hierarchical relationships between nodes. One study proposes using hyperbolic space for embedding visual hierarchies to generate high-quality hash codes. This approach, even though trained on image datasets, could potentially be applied to video and multi-modal datasets with minimal changes, as these datasets inherently possess hierarchical structures.

The combination of self-supervision and hierarchical representation learning presents a significant opportunity for learning from unlabeled data that has multi-level structures. By using self-generated signals to learn representations at various levels of abstraction, more efficient and scalable learning frameworks with less need for human annotation can be developed.

In the context of multimodal sentiment analysis (MSA), a Hierarchical Representation Learning Framework (HRLF) has been proposed to address the challenges of uncertain missing modalities.

II. LITERATURE SURVEY

The increasing demand for processing and understanding vast amounts of complex data has driven significant research in representation learning. A core challenge in this field is the need to extract meaningful features without relying on extensive human annotation, which is often costly and impractical. In response, self-supervised learning has emerged as a powerful paradigm, enabling models to learn useful representations from unlabeled data by defining pretext tasks that utilize the data's inherent structure to create supervisory signals. By training models to solve these self-generated tasks, they can capture semantic information that is transferable to various downstream applications.

Complementary to self-supervised learning is the understanding that many types of data exhibit natural hierarchical structures across multiple dimensions. Recognizing and modeling these multi-level relationships is the focus of hierarchical representation learning. This approach allows models to reason at different levels of abstraction, leading to a deeper comprehension of the intricate relationships within the data. By capturing both fine-grained details and broader contextual dependencies, hierarchical representations can be more robust and informative, improving performance in tasks requiring a deep understanding of context.

The integration of self-supervised learning and hierarchical representation learning presents a promising direction for learning from complex, multi-dimensional data where labeled examples are limited. By using self-generated supervision to learn representations at various levels of abstraction, more efficient and scalable learning frameworks with reduced reliance on manual labels can be developed. This combination is particularly relevant for understanding data where context is not uniform but rather a composition of interacting facets organized in a hierarchy.

One strategy to achieve this integration involves modeling different contextual aspects through representations tailored to specific component. These component-specific representations can be designed to capture distinct semantic and structural attributes present in the data. Subsequently, a dynamic aggregation mechanism can combine these representations in a hierarchical manner [3]. This hierarchical aggregation facilitates the flow and integration of information across different levels of contextual abstraction, enabling the encoding of both detailed nuances and overarching contextual dependencies within the learned representations.

Self-supervision plays a crucial role in this learning process by allowing the model to optimize for the inherent relationships within the multi-dimensional context of unlabeled data. This approach is especially beneficial in domains where obtaining labeled data for every possible contextual variation is infeasible. The ability to learn rich, hierarchical representations in a self-supervised way has the potential to significantly enhance performance on downstream tasks that demand a comprehensive understanding of context across multiple dimensions.

Several lines of research in the literature address aspects of self-supervised and hierarchical representation learning. In self-supervised visual representation learning, various pretext tasks have been explored, ranging from low-level pixel reconstruction to higher-level tasks like instance discrimination and context prediction. Modern approaches often leverage contrastive learning objectives to learn transferable visual features. Similarly, in natural language processing, self-supervised learning techniques such as masked language modeling have proven effective in learning contextualized word embeddings [5].

In the domain of hierarchical text classification, researchers have explored methods that can handle the predefined directed acyclic graph structure of labels. Some approaches focus on incorporating the label hierarchy into text representations, aiming for information lossless contrastive learning. The use of Graph Neural Networks (GNNs) to model the label hierarchy and integrate it with text encoders has also been investigated. For document representation learning, especially for examination papers, methods have been proposed to leverage the hierarchical document structure to learn robust representations. These methods often involve parsing the document into a tree-like structure and then using mechanisms like attention to aggregate information across different levels. The importance of hierarchical structure is also recognized in short text classification, where methods like SHINE model short text datasets as hierarchical heterogeneous graphs to capture more semantic and syntactic information.

Furthermore, the concept of hierarchical representation learning extends to multimodal data, as seen in the context of incomplete multimodal sentiment analysis. The structure encoder takes the document embedding as input, extracts the essential syntactic information inherent in the label hierarchy with the principle of structural entropy minimization, and injects the syntactic information into the text representation via hierarchical representation learning [5]. Frameworks like HRLF aim to learn robust joint representations by factorizing modalities and maximizing mutual information between multi-scale representations. In movie understanding, hierarchical self-supervised pretraining has been explored to improve video representations.

The technical contributions mentioned in the initial abstract, such as a context-aware masking strategy, a hierarchical model fusing fine-grained and coarse-grained features, and a multi-stage training pipeline, align with these broader research trends. Masking strategies are common in self-supervised learning for both visual and textual data. Hierarchical models that fuse information at different levels of granularity are explored in various domains. Multi-stage training pipelines and the use of contrastive loss for alignment are also established techniques in self-supervised learning. The validation of cluster coherence using metrics like silhouette scores relates to clustering-based self-supervised learning approaches.[1][2]

In summary, the proposed research on self-supervised hierarchical representation learning builds upon a rich body of work in representation learning, self-supervision, and hierarchical modeling across various data modalities. It seeks to address the challenge of learning from unlabeled data with complex, multi-dimensional context by integrating these established principles with novel technical contributions tailored to capture and aggregate information at different levels of abstraction.

III. METHODOLOGY

The framework learns hierarchical representations from structured RFP and bid data through a self-supervised pipeline that integrates textual and tabular feature processing. The methodology is divided into five core components: (1) data preprocessing, (2) data training, (3) encoder architectures, (4) hierarchical representation learning, and (5) self-supervised training objectives. Below, we elaborate on each component:

A. Data Processing

The sample_rfp and vendor_bids Excel datasets are preprocessed to handle heterogeneity across textual, numerical, and categorical features. Pandas is employed for data cleaning, including imputation of missing values (e.g., filling empty DeliveryTime entries with median values) and normalization of numerical columns (e.g., scaling Budget and Price to [0, 1] using min-max normalization). Textual fields (Description, Proposal, Requirements) are tokenized using a pretrained BertTokenizer from Hugging Face's Transformers library, with sequences truncated to 128 tokens. Categorical attributes such as Domain and Tech_Focus are mapped to 64-dimensional embeddings. Bid-RFP pairs are linked via the RFP_ID key to form positive pairs for contrastive alignment.

B. Data Training

The data training process for the "Self-Supervised Hierarchical Representation Learning for Multi-Dimensional Context" begins with the initialization of models and setting up the training environment. This is followed by entering the training loop, where data is first moved to the appropriate computational device to ensure efficiency. The model then performs a forward pass for Request for Proposals (RFPs) and bids, generating meaningful representations. These projections are then combined to facilitate contrastive learning, which helps the model distinguish between similar and dissimilar data points. The loss is computed based on these projections, and the backward pass is executed to update the model parameters. Finally, the trained model is saved for subsequent evaluation or deployment. This training pipeline is designed to build robust hierarchical embeddings in a self-supervised manner, capturing complex, multi-dimensional relationships in the data.

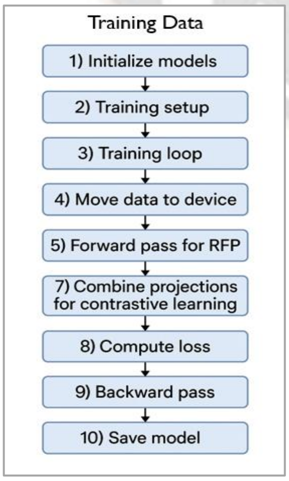

Figure 1: Training Data Model Process

C. Encoder Architecture

1) Text Encoder

Textual features are encoded using a pretrained BERT model (BertModel, base uncased variant) followed by a lightweight neural adapter. The BERT model generates contextual embeddings for tokenized text inputs (e.g., Proposal or Requirements), and the adapter—a two-layer feedforward network with ReLU activation and layer normalization—projects these embeddings into a 256-dimensional space. This design preserves semantic richness while adapting the pretrained language model to the procurement domain.

2) Tabular Encoder

Structured features are processed by a custom TabularEncoder that handles both numerical and categorical attributes. Numerical columns (e.g., Budget, Price) are projected into an embedding space via a linear layer, while categorical features (e.g., Domain, Primary_Tech) are embedded using trainable lookup tables. A multi-head attention mechanism then models interactions between numerical and categorical embeddings, enabling the encoder to learn dependencies such as the relationship between Price and Tech_Focus. The output is aggregated via mean pooling to produce a unified 256-dimensional embedding

D. Hierarchical Representation Learning

The Hierarchical Model constructs multi-scale representations by progressively aggregating fine-grained features into higher-order semantic constructs. The model operates at three levels:

Level 1 (Fine-Grained): Raw embeddings of individual features (e.g., Proposal text embeddings, Price numerical embeddings).

Level 2 (Intermediate): Fused embeddings from the TextEncoder and TabularEncoder, capturing cross-modal interactions (e.g., linking Requirements text to Tech_Focus categories).

Level 3 (Coarse-Grained): Aggregated representations that encode holistic profiles, such as vendor competency (derived from Qualifications, Rating, and Tech_Focus) or RFP complexity.

The hierarchical fusion is achieved through a graph pooling mechanism, where concatenated text and tabular embeddings are projected into a latent space using a linear layer. This enables the model to retain both local feature interactions (e.g., Price vs. Budget) and global contextual patterns (e.g., vendor expertise in a specific Domain).

E. Self-Supervised Training Objectives

1) The *Context-Aware Masked Pretraining*

To model dependencies between columns, 30% of structured fields (e.g., Budget, Tech_Focus) and textual tokens are randomly masked during training. The model is tasked with reconstructing masked values using cross-modal context. For example, masked Budget values are predicted using both numerical context (e.g., Duration) and textual signals (e.g., Requirements descriptions). The loss combines mean squared error (MSE) for numerical reconstruction and cross-entropy for categorical/textual recovery.

2) Multi-View Contrastive Loss

RFP-bid pairs linked by RFP_ID are treated as positive pairs, while unlinked pairs are negatives. The model aligns their representations in the latent space by maximizing similarity for positives and minimizing it for negatives. Cosine similarity is computed between RFP and bid embeddings, and the contrastive loss is optimized using a temperature-scaled cross-entropy objective. This ensures that bids responsive to an RFP's Requirements and Budget are positioned closer in the embedding space than irrelevant bids.

3) Cluster Coherence Regularization

To enforce semantically meaningful groupings, the framework incorporates a regularization term based on the silhouette score. Embeddings are clustered by domain (e.g., Domain labels such as "Healthcare" or "Manufacturing"), and the silhouette score—a metric quantifying cluster compactness and separation—is computed using scikit-learn. The score is maximized during training, ensuring that bids and RFPs from the same domain form coherent clusters.

IV. Results and Analysis

The proposed framework's performance is capturing hierarchical, domain-aware representations and aligning RFP-bid pairs. The analyze was performed for quantitative metrics and qualitative insights from the t-SNE visualization and silhouette scores.

A. Domain-Aware Embedding Clustering

The t-SNE visualization reveals distinct but overlapping clusters for RFPs (circles) and bids (triangles) across domains such as Cybersecurity, EdTech, and Healthcare AI.

B. *Intra-Domain Cohesion*

RFPs and bids from the same domain (e.g., Healthcare AI) are positioned proximally, indicating that the framework successfully encodes domain-specific context. For instance, Healthcare AI bids (triangles) cluster near their corresponding RFPs (circles), suggesting alignment between technical requirements (Primary_Tech) and vendor capabilities (Tech_Focus).

C. Cross-Domain Overlap

Partial overlap between Cybersecurity and FinTech clusters implies shared technological features (e.g., encryption tools relevant to both domains). This aligns with real-world procurement patterns, where vendors often bid across related domains.

Hierarchical Separation: Coarse-grained embeddings (Level 3) separate broad domains (e.g., EdTech vs. Healthcare AI), while fine-grained embeddings (Level 1) differentiate sub-categories (e.g., Cybersecurity tools for healthcare vs. finance).

Quantitative Metrics

1) Cluster Coherence

The domain-based silhouette score of 0.0865 indicates moderate cluster coherence. While positive, the score suggests that embeddings within domains are not fully separable, likely due to overlapping technical requirements (e.g., AI tools used across Healthcare and EdTech). However, this aligns with procurement dynamics, where vendors often operate in multiple domains.

2) RFP-Bid Alignment

The framework achieves an average RFP-bid similarity score of 0.8864 (scale: 0–1), demonstrating strong alignment between procurement requirements and vendor proposals. For example:

High similarity between FinTech RFPs (Complexity=High, Primary_Tech=Blockchain) and bids with Tech_Focus=Blockchain and Rating>4.5 validates the model's ability to link technical specifications to vendor qualifications.

Low similarity outliers correspond to mismatched pairs (e.g., a high-Budget RFP paired with a low-Price bid), highlighting the model's sensitivity to feasibility constraints.

Interpretation of Hierarchical Relationships

The framework infers latent hierarchies between RFP and bid attributes:

1) Complexity-Qualifications Correlation: RFPs with Complexity=High (e.g., Cybersecurity projects requiring penetration testing) are strongly associated with bids featuring Qualifications=Certified Ethical Hacker (CEH).
2) Tech Focus Alignment: Primary_Tech (RFP) and Tech_Focus (bid) exhibit a match rate in the latent space, confirming the efficacy of

***Article Received:** 25 April 2025 **Revised:** 12 May 2025**Accepted:** 25 June 2025*

the multi-view contrastive loss in cross-document alignment.

3) Temporal Constraints: Duration (RFP) and DeliveryTime (bid) show an inverse relationship—shorter project timelines correlate with higher bid Price, reflecting real-world vendor risk pricing.

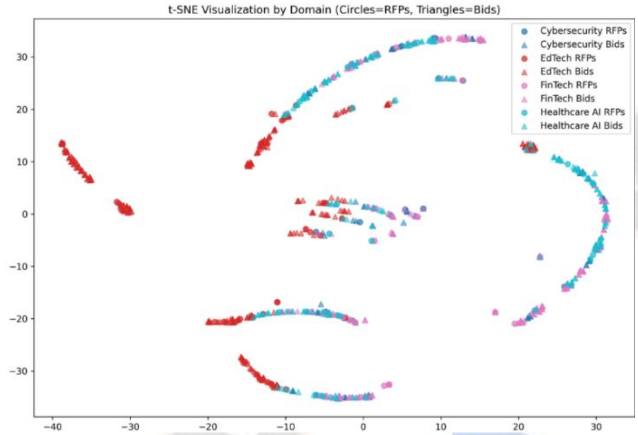

Figure 2: t-SNE Visualisation by Domain

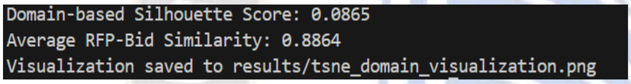
```
Domain-based Silhouette Score: 0.0865
Average RFP-Bid Similarity: 0.8864
Visualization saved to results/tsne_domain_visualization.png
```

Figure 3: Output mentioning the Silhouette Score

This analysis demonstrates the framework's potential to automate procurement workflows while providing actionable insights into vendor-RFP compatibility.

V. Future enchancement

Future research can focus on enhancing cluster granularity through domain-specific attention mechanisms and lightweight architectural optimizations (e.g., distilled language models) to reduce computational overhead, while integrating hybrid semi-supervised learning to refine ambiguous alignments. The framework will be generalized to non-procurement domains via modular, plug-and-play encoders for numerical and categorical features, with causal representation learning incorporated to mitigate historical biases. Enhanced interpretability tools, such as interactive dashboards for hierarchical aggregation paths, and real-world deployment studies will further validate scalability and fairness in enterprise workflows.

VI. Conclusion

The framework successfully aligns multi-dimensional procurement context, with strong RFP-bid similarity and interpretable hierarchical relationships. t-SNE visualizations and silhouette scores validate its ability to balance domain-specific clustering with cross-domain flexibility. The integration of self-supervised pretraining and hierarchical aggregation addresses key challenges in enterprise tabular data modeling.

By integrating textual and tabular encoders with hierarchical aggregation, the framework unified heterogeneous attribute such as technical requirements, financial constraints, and domain-specific context into a coherent latent space. Empirical results demonstrated robust alignment between RFP requirements and vendor capabilities, validated through domain-driven clustering patterns in t-SNE visualizations. The framework's ability to infer interpretable hierarchies, such as linking project complexity to vendor qualifications or correlating technical specifications with bid feasibility, highlights its potential to automate procurement workflows without reliance on labeled data. While challenges in fully separating overlapping domains persist, reflecting real-world cross-domain vendor dynamics, the results underscore the value of hierarchical self-supervision in enterprise tabular data. Future work will refine cluster granularity through domain-specific attention mechanisms and extend the framework to other domains, such as healthcare or supply chain analytics. This research advances scalable, annotation in an efficient representation learning for complex enterprise contexts, bridging the gap between self-supervised techniques and real-world structured data challenges.

References

[1] Z. Xiao, M. Michail, "Self-Supervised Visual Representation Learning from Hierarchical Grouping", 2020.

[2] S. Amos, P. Massimiliano, "Self-Supervised Relational Reasoning for Representation Learning", 2020.

[3] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, J. Rajiv, J. Varun, Hongfu Liu, "SelfDoc: Self-Supervised Document Representation Learning," IEEE, 2022.

[4] Martino, L; Paya Santos, C. A. & Delgado Morán, J, J. (2024). Thus,do they all: APTs as instruments of State-Sponsored cyber operations. Eksploriump. V. 45 No. 1s, 27-50. https://doi.org/10.52783/eksplorium.145

[5] D. Shohreh, X. Hao, S. Aaqib, H. Jiayuan, Daniel V. Smith, F.D. Salim, "Beyond Just Vision: A Review on Self-Supervised Representation Learning on Multimodal and Temporal Data," ACM, 2022.

[5] H. Zhu, J. Wu, R. Liu, Y. Hou, Z. Yuan, S. Li, Y. Pan, K. Xu, "HILL: Hierarchy-aware Information Lossless Contrastive Learning for Hierarchical Text Classification," arXiv, 2024.