Article

Enhancing Sentence Representation with Syntactically Structural Transformer

Petrenko Elizabeth * , Rodolfo Patel , Fomuso Jose

Posted Date: 11 December 2023

doi: 10.20944/preprints202312.0600.v1

Keywords: Syntactic tree; Transformer; Self-attention

Article

Enhancing Sentence Representation with Syntactically Structural Transformer

Petrenko Elizabeth *, Rodolfo Patel and Fomuso Jose

Briar Cliff University
* Correspondence: elizabeth@briarcliff.edu

Abstract: The evolution of sentence representation learning, especially with parse tree encoders, has shown remarkable progress. Traditional approaches predominantly rely on recursive encoding of tree structures, which impedes parallel processing capabilities. Additionally, these methods often overlook the significance of dependency tree arc labels. To overcome these limitations, we introduce the Syntax-Enhanced Transformer (SET), incorporating a novel dual-attention mechanism that integrates relation-focused attention alongside traditional self-attention. This design effectively encodes both dependency and spatial positional relationships within sentence dependency trees. Our approach innovatively incorporates syntactic information into the Transformer framework without compromising its inherent parallelizability. The SET demonstrates superior or comparable performance to contemporary methods across various sentence representation tasks, significantly enhancing computational efficiency.

Keywords: syntactic tree; transformer; self-attention

1. Introduction

The use of distributed sentence representations has become integral in numerous natural language processing (NLP) tasks. Various operators, including recurrent neural networks (RNN) [1,2], convolutional neural networks (CNN) [3–5], recursive convolutional neural networks (RCNN) [8,9], and Transformers [10,11], have been employed for mapping lexical representations into cohesive sentence representations. While effective, these models often neglect the rich syntactic information inherent in sentence structures [12–20]. Contrary to earlier sequence-structure encoding methods, tree-based models encode sentences by recursively processing tree structures through diverse compositional functions. Parse trees, illustrating syntactic sentence structures, are repositories of grammatical nuances. Recursive models have been developed to encode sentences and their parse trees, leveraging a variety of compositional functions in a bottom-up fashion. Despite their innovative approach, RNN-based models suffer from limited parallelization capabilities, exacerbated by the heterogeneous nature of tree structure inputs, hindering batch processing and parallel training.

In response, the Transformer model has gained widespread popularity, attributed to its exceptional parallelism and performance. However, existing Transformer adaptations, such as Tree-Transformer [25], still adhere to a recursive mechanism, limiting parallel training. Furthermore, dependency trees encapsulate both topological structures and dependency relation types. These labels are pivotal, yet seldom effectively distinguished in existing models. This gap prompts an inquiry: Can we harness syntax tree information [31] extensively without compromising the Transformer's parallel processing feature?

Addressing this, we present the Syntax-Enhanced Transformer (SET), with a relation-attention mechanism, providing an affirmative response to the aforementioned challenge. The SET's dual-attention mechanism seamlessly blends dependency information with the self-attention mechanism. It decomposes trees into word relations, assigning learnable vectors to dependency relations and relative tree positions. These vectors are pooled into scores and amalgamated with self-attention scores during parallel word pair processing. Additionally, a gating mechanism correlates these relation-attention scores with sentence semantics, dynamically adjusting their contribution to the self-attention scores based on lexical context.

Our model's effectiveness is validated on four benchmark sentence representation datasets, where it surpasses or equals existing methodologies, notably on SICK-E and SICK-R, and competes closely on SST-2 and MRPC. The SET not only elevates performance but also achieves a 6-8 fold increase in training and testing speed compared to recursive models. Our comprehensive studies, including ablations and case analyses, further substantiate the efficiency and validity of our proposed mechanism.

Our primary contributions are as follows:

- Introducing the Syntax-Enhanced Transformer and its dual-attention mechanism, which preserves the Transformer's parallelization advantages while optimizing dependency tree utilization.
- Demonstrating our model's proficiency on four sentence representation tasks: SICK-E, SICK-R, SST-2, and MRPC, achieving state-of-the-art results on two and competitive performance on the others. The effectiveness of our approach is further supported by detailed ablation and case studies.

2. Proposed Model: Syntax-Enhanced Transformer (SET)

The Syntax-Enhanced Transformer (SET) structure is grounded in a robust three-layer Transformer encoder. We consider an input sequence $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{l \times d}$. To align with the syntactic tree structure, a unique $[root]$ token, symbolizing the $ROOT$ node of the parse tree, is prepended to the sentence, resulting in the modified input $X = [root, x_1, x_2, ..., x_n]$. The output from the model is the transformed sequence $Y' = [y_{root}, y_0, y_1, .., y_n] \in \mathbb{R}^{(l+1) \times d}$, with the $ROOT$ node's output serving as the holistic sentence representation.

In the conventional Transformer setup, the input comprises the sum of word embeddings and positional vectors. These components respectively capture the training corpus' word distribution and the positional relevance within a sentence. The SET introduces an innovative layer, the level embeddings, delineating the hierarchical levels of words within the parse tree (measured as the distance from the root node). These level embeddings, akin to positional vectors, are set as learnable vectors for distinct tree levels, enriching the input to our model with nuanced syntactic depth.

2.1. High-order Dependency Relation Enhancement

In the Transformer's self-attention mechanism, attention scores are the scaled-dot products of the words' hidden states, reflecting lexical semantic relationships. Similarly, the relational dynamics within a dependency parse tree can be interpreted as correlations among words. To encapsulate this, we integrate dependency relations into the self-attention scores. This integration is poised to capture the sentence's syntactic structure.

Previous approaches have underutilized relation types in dependency trees. Addressing this, we assign a unique learnable vector to each relation type. These vectors are processed through different linear layers within each attention head, contributing to the calculation of augmented attention scores. These enhanced scores are then added to the conventional attention scores as a bias. The relation-attention score for a pair of words ($word_i$ attending to $word_j$) in a given attention head is computed as:

$$S^r_{ij} = r_{ij} V_r; , \tag{1}$$

where $r_{ij} \in \mathbb{R}^{d_r}$ denotes the relation vector from $word_i$ to $word_j$, and $V_r \in \mathbb{R}^{d_r \times 1}$ represents the learnable vector in the current attention head.

2.2. Semantic Learning via Syntax

The relation-attention and self-attention scores are initially independent, potentially limiting the model's adaptability. To ensure semantic relevance, we introduce a semantic gated mechanism. This mechanism dynamically adjusts relation scores based on word semantics, preventing inappropriate influence on self-attention. We compute gated scores using relation embeddings and the key word's hidden states, applying a tanh activation to scale the results within the range $g \in [-1, 1]$. The attention score is reformulated as follows:

$$S_{ij} = (1 - g_{ij}) \odot S^e_{ij} + g_{ij} \odot S^r_{ij};, \tag{2}$$

$$g_{ij} = sigmoid((h_i W_{g,e} + r_{ij} W_{g,r}) V_g);, \tag{3}$$

where S^e_{ij} represents the self-attention score between words. $W_{g,r} \in \mathbb{R}^{d_r \times d_r}$, $W_{g,e} \in \mathbb{R}^{d_e \times d_r}$, and $V_g \in \mathbb{R}^{d_r \times 1}$ are independent trainable matrices and vectors in each attention head. h_i signifies the hidden states of $word_i$, which in the first layer is the sum of word embeddings, position embeddings, and level embeddings: $h^0_k = e_k + p_k + l_k$, and in subsequent layers, it is the output from the preceding layer.

2.3. Structural Modeling

To avoid sparsity in the relation matrix, which arises from considering only adjacent nodes in the tree structure, we enrich the matrix with additional dependencies. Specifically, we identify non-adjacent word pairs within a certain threshold distance and treat them as extended dependency relations. We utilize a relative position encoding of the shortest paths to represent relations between these non-adjacent nodes. The construction of the relation matrix R (with relation vector r_{ij} corresponding to the i-th row and j-th column of R) is illustrated in Figure 2.

In half of the attention heads, we apply a subtree mask, allowing query words to attend only to their descendants in the dependency tree. This focused attention emphasizes grammatical dependencies and simulates a bottom-up accumulation process, enhancing the model's syntactic analysis capabilities.

3. Experiments

This section presents a comprehensive evaluation of our SET model, including dataset descriptions, experimental setup, results analysis, and an in-depth ablation study.

Table 1. Comparative Evaluation Results on Four Datasets.

Models	**SICK-R**	**SICK-E**	**SST-2**	**MRPC**
	(MSE)	(Acc.)(%)	(Acc.)(%)	(Acc.)(%)
LSTM [12]	.2831	76.80	84.90	71.70
BiLSTM [12]	.2736	82.11	87.50	72.70
RNTN [16]	-	59.42	85.40	66.91
DT-RNN [16]	.3848	63.38	86.60	67.51
Tree-LSTM$_{DT}$ [12]	.2734	82.00	85.70	72.07
Tree-LSTM$_{CT}$ [12]	.2532	83.11	88.00	70.07
BiTree-LSTM [32]	.2736	-	90.30	-
TagHyperTreeLSTM [20]	-	83.90	**91.20**	-
USE [33]	-	81.15	85.38	**74.96**
StructTransformer [4]	-	82.30	87.80	-
Tree-Transformer$_{DT}$ [25]	.2774	82.95	83.12	70.34
Tree-Transformer$_{CT}$ [25]	.3012	82.72	86.66	71.73
SET (Ours)	**.2634**	**87.50**	90.06	75.31

3.1. Setups

The SET model is rigorously tested on diverse sentence representation tasks: text classification (Stanford Sentiment Treebank - SST-2), text semantic matching (SICK dataset), and paraphrase detection (MRPC). We have utilized existing datasets with established benchmarks to ensure a comprehensive and fair evaluation.

Baseline Comparison: Our model is benchmarked against a suite of state-of-the-art models encompassing various architectures, including LSTM, Bi-LSTM, several tree-based models, and Transformer variants. This comparison aims to highlight the advancements the SET model brings to sentence representation tasks.

Experimental Details: The experiments were conducted using the Stanford dependency parser for sentence parsing. Word embeddings were initialized with GloVe vectors and updated during training. The SET model, with its unique self-attention and relation-attention mechanisms, was optimized using AdaGrad with specific hyperparameters. The evaluation encompassed various metrics, including accuracy and Pearson correlation, depending on the task.

3.2. Results

The SET model demonstrates superior performance on the SICK-E and SICK-R datasets, outstripping leading models by a significant margin. In the SST-2 task, the SET achieves comparable results to the top-performing models and shows clear advantages over other Transformer-based models. In the MRPC task, the SET ranks second, demonstrating its robustness across different sentence representation tasks.

The efficiency of the SET model is highlighted in Table 2, showcasing its significantly reduced training and testing times compared to the Tree-LSTM model. This efficiency is attributed to the SET's advanced parallelization capabilities.

Table 2. Training and Testing Efficiency Comparison between SET and Tree-LSTM.

Model	Testing	Training
Tree-LSTM	8s	282s
SET (Ours)	<1s	57s

3.3. Ablation Study

An ablation study was conducted to dissect the contributions of different components of the SET model. The study compared the complete SET model with variants lacking specific features, such as level embedding and gating mechanism. This comparison reveals the individual and combined impacts of these components on the model's performance, thereby validating the effectiveness and rationality of the SET's unique design.

Table 3. Ablation Study Results for the Syntax-Enhanced Transformer.

Models	**SICK-R**	**SICK-E**	**SST-2**	**MRPC**
	(MSE)	(Acc.)(%)	(Acc.)(%)	(Acc.)(%)
Transformer	0.2833	83.34	87.19	72.41
SET$_{lr}$	0.2545	84.76	88.21	73.01
SET$_{rg}$	0.2526	84.93	88.53	73.11
SET$_{full}$	0.2428	85.13	88.77	73.58

4. Conclusion

In this work, we introduce the Syntax-Enhanced Transformer (SET), a novel approach that innovatively incorporates dependency tree data into the self-attention mechanism of the Transformer. The SET model is characterized by its relation-attention mechanism, which is adeptly designed to integrate and leverage syntactic structures within sentences. Further distinguishing this model is the incorporation of a unique gating mechanism. This mechanism is skillfully crafted to create a synergy between syntactic relationships and semantic context, ensuring that the syntactic structure enriches the semantic understanding in a meaningful way. One of the most compelling features of the SET model is its ability to maintain the inherent parallel processing capabilities of the traditional Transformer architecture. This aspect is particularly significant as it allows the model to handle large datasets and complex computations efficiently. Moreover, the SET model demonstrates remarkable enhancements in its handling of syntactic information, a feature that sets it apart from its predecessors. Our extensive evaluations across a range of sentence representation tasks highlight the SET model's superior performance. Notably, when compared against a variety of baseline models, the SET consistently shows improvements not only in efficiency but also in overall performance. This improvement is a testament to the effectiveness of integrating syntactic structure into the self-attention mechanism and the benefits of the added gating mechanism. The SET thus stands out as a significant advancement in the field of natural language processing. It opens new avenues for exploring and integrating syntactic structures in deep learning models, paving the way for more nuanced and context-aware language understanding systems. The findings from our research underscore the potential of combining syntactic and semantic information in language models, promising exciting developments for future research in this domain.

References

1. Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.
2. Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.
3. Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.
4. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
5. Yoon Kim. Convolutionalneuralnetworksforsentence classification.
6. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
7. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
8. Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
9. Han Zhao, Zhengdong Lu, and Pascal Poupart. Self-adaptive hierarchical sentence model. *arXiv preprint arXiv:1504.05070*, 2015.
10. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
11. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
12. Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

13. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
14. Richard Socher, Cliff Chiung-Yu Lin, Andrew Y Ng, and Christopher D Manning. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, 2011.
15. Hao Fei, Yafeng Ren, and Donghong Ji. Mimic and conquer: Heterogeneous tree structure distillation for syntactic NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 183–193, 2020.
16. Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.
17. Hao Fei, Yafeng Ren, and Donghong Ji. Improving text understanding via deep syntax-semantics communication. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 84–93, 2020.
18. Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.
19. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
20. Chunlin Xu, Hui Wang, Shengli Wu, and Zhiwei Lin. Treelstm with tag-aware hypernetwork for sentence representation. *Neurocomputing*, 434:11–20, 2021.
21. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
22. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
23. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
24. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
25. Mahtab Ahmed, Muhammad Rifayat Samee, and Robert E Mercer. You only need attention to traverse trees. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 316–322, 2019.
26. Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. Semi-supervised text style transfer: Cross projection in latent space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4939–4948, 2019.
27. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
28. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12794–12802, 2021.
29. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
30. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
31. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 549–559, 2021.
32. Zhiyang Teng and Yue Zhang. Head-lexicalized bidirectional tree lstms. *Transactions of the Association for Computational Linguistics*, 5:163–177, 2017.

33. Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.