Multi-Granularity Reasoning for Natural Language Inference

Chunling Xi[1], Di Liang[2*]
[1]Pacific Insurance Technology Co., Ltd. [2]Lixin Information Services (Shenzhen) Co., Ltd.
xichunling987@163.com

***Abstract*—Natural Language Inference (NLI) is a fundamental task in natural language understanding that requires determining the logical relationship between a premise and a hypothesis. Despite the remarkable success of transformer-based pre-trained models, most existing approaches primarily rely on the final-layer token representations, which are often insufficient for capturing the complex and hierarchical semantic interactions required for effective reasoning. In particular, fine-grained lexical cues, phrasal compositions, and higher-level contextual semantics are typically entangled or diluted in a single representation space. To address these limitations, we propose a novel *Multi-Granularity Reasoning Network* (MGRN) that explicitly leverages hierarchical semantic features within an interactive reasoning space. The proposed framework mimics the human cognitive process of language understanding, which naturally progresses from shallow lexical matching to deeper semantic abstraction and logical reasoning. By integrating semantic information across multiple granularities in a progressive and structured manner, MGRN is able to uncover intricate semantic relationships underlying natural language expressions. Extensive experiments on multiple public benchmarks demonstrate that MGRN consistently outperforms strong baseline models, validating the effectiveness and robustness of the proposed approach.**

***Index Terms*—Multi-Granularity Reasoning, Semantic Interaction, Natural Language Inference**

I. Introduction

Natural Language Inference (NLI), also known as Recognizing Textual Entailment, constitutes a fundamental task in natural language understanding that requires determining the logical relationship between a premise and a hypothesis. The three possible relationships are "entailment" (the premise logically supports the hypothesis), "contradiction" (the premise contradicts the hypothesis), or "neutral" (no conclusive relationship can be established). As a cornerstone of natural language understanding, NLI presents substantial challenges due to the inherent complexity of natural language, which includes diverse expressions, rich semantic nuances, and intricate contextual dependencies [2]. For instance, accurately discerning such relationships often demands not only lexical and syntactic awareness but also commonsense knowledge and multi-level reasoning.

Given its remarkable success, we hypothesize that more structured and multi-faceted attention signals can further enhance a model's ability to understand language and perform natural language inference. Conventional attention mechanisms capture pairwise word-level alignments between sentences. The advent of multi-head attention [12], [32], [33], [42], [42] extended this by enabling the model to jointly attend to information from different representation subspaces, effectively capturing various types of relationships. In this study, we build upon this idea and introduce a novel interactive tensor structure that captures higher-order semantics, not only between words but also among phrases and broader contextual elements. This structure enhances the model's ability to handle complex linguistic phenomena such as paraphrasing and lexical variation.

To leverage these richer interaction patterns effectively, we propose a novel **Multi-Granularity Reasoning Network (MGRN)**. This framework is specifically designed to aggregate and reason over multi-level semantic features through a progressive hierarchical process, simulating the human cognitive progression from surface-level understanding to deep semantic comprehension. We extensively evaluate MGRN on widely used standard NLI benchmarks (SNLI and MultiNLI) and observe consistent and notable improvements over strong competitive baselines. To further assess and verify its general applicability, we adapt MGRN to the paraphrase identification task by explicitly framing it as a binary NLI classification task (paraphrase as entailment, non-paraphrase as neutral). Comprehensive experiments on the Quora Question Pairs dataset and seven additional diverse benchmarks demonstrate the robustness and effectiveness of the proposed approach.

The main contributions of this work are summarized more concretely as follows: Firstly, we propose a powerful Multi-Granularity Reasoning Network (MGRN) that integrates hierarchical semantic features using a carefully structured interactive tensor, significantly enhancing overall inference accuracy. Secondly, the proposed model effectively captures high-order semantic interactions between sentence pairs, extending far beyond traditional alignment methods to incorporate complex compositional and meaningful phrasal relationships. Finally, extensive and rigorous experiments on multiple widely recognized public datasets show that MGRN consistently and clearly outperforms several competitive baselines, demonstrating its effectiveness, robustness, and general applicability across various challenging NLP tasks.

II. Related Work

Neural Language Inference Early studies on NLI primarily relied on hand-crafted features, logical rules, and syntactic or semantic parsing [1]. These approaches were limited by feature engineering and scalability. The introduction of large-scale annotated datasets such as SNLI [2] and later

MultiNLI enabled data-driven neural models to dominate the field. Early neural architectures focused on sentence encoding, where each sentence was independently mapped into a fixed-length vector using CNNs or RNNs, followed by a classifier [6]. To address this limitation, interaction-based models were proposed to explicitly model cross-sentence alignments at the word or phrase level [7], [34]–[37]. Attention mechanisms [8], [38]–[40] further improved performance by enabling soft alignment and comparison between sentence representations. Subsequent work explored deeper and more expressive architectures. Residual connections [3] and dense interaction layers enabled multi-step inference and iterative reasoning. Some approaches incorporated syntactic or semantic structures, such as dependency trees or semantic graphs, to guide reasoning [1]. Despite these advances, most neural NLI models still rely heavily on single-granularity interactions, typically focusing on token-level alignments, which limits their ability to capture higher-order semantic composition and abstract reasoning.

Pre-trained Language Model Methods Pre-trained language models have significantly reshaped NLI research by providing powerful contextualized representations learned from large-scale corpora. BERT [9] demonstrated that bidirectional self-attention and masked language modeling objectives yield representations that transfer effectively to NLI. Subsequent models such as RoBERTa [10], XLNet [11], [26]–[31], [44], [52], and CharBERT [17] further improved performance through refined pre-training strategies, architectural modifications, or enhanced subword modeling. Beyond standard fine-tuning, several studies have explored explicit cross-sentence modeling within pre-trained frameworks. For example, joint encoding and interaction mechanisms [4], [12], [22] integrate alignment and comparison operations to enhance inference capability. Other works incorporate external linguistic or semantic knowledge [15], syntactic features, or handcrafted signals [23] to compensate for data sparsity and improve interpretability. However, despite their strong performance, pre-trained models typically compress linguistic information into the final-layer representations used for downstream classification. Recent analyses suggest that different layers encode distinct types of linguistic knowledge, ranging from surface features to abstract semantics. Relying solely on final-layer representations may obscure useful intermediate semantic signals, particularly for cases requiring fine-grained reasoning or multi-step inference. This observation motivates approaches that explicitly exploit representations at multiple levels of abstraction.

Robustness Evaluation Although modern NLI models achieve high accuracy on standard benchmarks, numerous studies have shown that they are brittle under small perturbations, distribution shifts, or adversarial attacks [5], [41], [43], [45], [46]. Models may rely on shallow heuristics or spurious correlations rather than genuine semantic understanding, leading to degraded performance in challenging or out-of-distribution scenarios. To address these issues, robustness-oriented evaluation frameworks have been proposed, including adversarial test sets, stress tests targeting specific linguistic phenomena, and systematic data augmentation toolkits such as TextFlint [24], [50], [51], [53]. Diagnostic platforms like Explainaboard [25], [47]–[49] emphasize fine-grained analysis across different capabilities. These efforts consistently reveal a substantial gap between benchmark performance and true reasoning ability, highlighting the importance of models that can integrate and reason over semantic information at multiple granularities. Our work aligns with this direction by explicitly modeling hierarchical semantic interactions to improve both performance and robustness.

III. Method

We define the natural language inference as a classification task that predicts the relation y $\in$ Y for a given pair of sentences.

A. Input preprocessing

Given two texts $S_1 = \{x_1^1, x_2^1, \ldots, x_n^1\}$ and $S_2 = \{x_1^2, x_2^2, \ldots, x_m^2\}$, we add special tags [CLS] and [SEP] before and after them to adapt to the input form of BERT, that is, [CLS], $x_1^1, x_2^1, \ldots, x_n^1$, [SEP], $x_1^2, x_2^2, \ldots, x_m^2$, [SEP]. For BERT For example, the input needs to be obtained by adding the embedding vectors of three parts:

$$\mathbf{E}_i = \mathbf{T}_i + \mathbf{S}_i + \mathbf{P}_i, \tag{1}$$

$\mathbf{T}_i$, $\mathbf{S}_i$, $\mathbf{P}_i$ respectively represent the Token Embedding, Segment Embedding and Position Embedding of the ith word. Here, Segment Embedding distinguishes the first and second paragraphs of text (i.e., sentence A/B tags).

B. Multi-layer Transformer representation

BERT is composed of L layers of Transformer Blocks stacked together, each layer of which contains Multi-Head Self-Attention and Feed-Forward Network. If $\mathbf{H}^{(0)} = [\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_{N'}]$ represents the input embedding sequence (where $N' = n + m + 3$, including [CLS] and 2 [SEP]), then the output of the lth layer can be recorded as

$$\mathbf{H}^{(l)} = \text{TransformerBlock}_l(\mathbf{H}^{(l-1)}), \quad l = 1, 2, \ldots, L. \tag{2}$$

The tensor of the first paragraph of text (removing [CLS], [SEP]) in $\mathbf{H}^{(l)}$ is recorded as $\mathbf{H}_1^{(l)} \in \mathbb{R}^{n \times d}$, and the tensor corresponding to the second paragraph of text is recorded as $\mathbf{H}_2^{(l)} \in \mathbb{R}^{m \times d}$. Where d is the hidden dimension of BERT.

C. Construction of the interaction matrix between layers

Interaction matrix obtained by bitwise multiplication In order to allow the model to explicitly capture the interaction between the first and second paragraphs of text in terms of features, element-wise multiplication is performed on the two sentence representations of the same layer. Specifically, for the lth layer, if we set $\mathbf{h}_1^{(l,i)} \in \mathbb{R}^d \quad (1 \le i \le n)$, $\mathbf{h}_2^{(l,j)} \in \mathbb{R}^d \quad (1 \le j \le m)$,, then the interaction tensor (or interaction matrix) is defined as:

$$\mathbf{M}_{i,j}^{(l)} = \mathbf{h}_1^{(l,i)} \odot \mathbf{h}_2^{(l,j)} \in \mathbb{R}^d, \quad \forall\, i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{3}$$

Where "$\odot$" represents the multiplication operation of the corresponding position. Thus, $\mathbf{M}^{(l)} \in \mathbb{R}^{n \times m \times d}$ is obtained,

which explicitly encodes the word-by-word interaction information of two sentences in the lth layer representation space.

D. Multi-layer stacking

The interaction matrices of each layer obtained by the above formula (3) are concatenated or stacked on the layer dimension to obtain a complete multi-layer interaction representation:

$$\overline{\mathbf{M}} = \left[\mathbf{M}^{(1)}; \mathbf{M}^{(2)}; \ldots; \mathbf{M}^{(L)}\right] \in \mathbb{R}^{n \times m \times d \times L}, \quad (4)$$

Where "$[\cdot;\cdot]$" means stacking on the new layer dimension. $\overline{\mathbf{M}}$ aggregates the learning results of sentence interaction features from the first layer to the Lth layer, which can provide richer and fine-grained feature information for the subsequent network (DenseNet).

E. DenseNet feature extraction

DenseNet structure and input DenseNet (Densely Connected Convolutional Network) was first proposed for image processing, achieving efficient gradient propagation and feature reuse by concatenating features between layers. The outputs of all previous layers are concatenated as inputs to subsequent layers; if the kth layer output is $\mathbf{z}_k$, a typical dense connection in a Dense Block is:

$$\mathbf{z}_k = H_k([\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{k-1}]),$$

where $H_k(\cdot)$ denotes the nonlinear transformation of the kth layer, and $[\cdot]$ represents feature concatenation.

$$\mathbf{Z}_m = \text{Concat}\big(\mathbf{Z}_0, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{m-1}\big), \quad (5)$$

Where $\mathbf{Z}_0$ is the input tensor of the current Block (which can be regarded as $\overline{\mathbf{M}}$), $\mathbf{Z}_m$ is the final output of the Block, and $\text{Concat}(\cdot)$ represents the concatenation operation.

DenseNet output representation The multi-layer interaction matrix $\overline{\mathbf{M}}$ passes through several Dense Blocks (and necessary transition layers, Transition Layer) in sequence to obtain the final feature tensor (or vector):

$$\mathbf{F} = \text{DenseNet}\big(\overline{\mathbf{M}}\big). \quad (6)$$

Where $\mathbf{F}$ is the high-level semantic representation of the sentence pair after multi-step nonlinear mapping.

F. Classification layer

Finally, a fully connected network is added to the output representation $\mathbf{F}$ of DenseNet and the softmax function is used to obtain the probability distribution of the three-classification. If the three categories are $\{C_1, C_2, C_3\}$, it can be written as

$$\mathbf{p} = \text{softmax}\big(W\mathbf{F} + \mathbf{b}\big), \quad (7)$$

where $W \in \mathbb{R}^{3 \times \dim(\mathbf{F})}$, $\mathbf{b} \in \mathbb{R}^3$. During inference, the predicted category is

$$\hat{y} = \arg\max_i \, p_i. \quad (8)$$

This end-to-end architecture enables the model to jointly leverage multi-layer semantic representations and fine-grained interaction features, allowing effective reasoning across multiple granularities for natural language inference.

IV. Experiments Setting

Datasets and baselines We conduct the experiments to test the performance of CIRN on 10 large-scale publicly benchmark datasets(GLUE benchmark,MRPC, QQP, STS-B, MNLI, RTE, and QNLI). And to evaluate the effectiveness of our proposed CIRN, we mainly introduce BERT [9], SemBERT [15], SyntaxBERT, UERBERT [23] and multiple other PLMs [9] for comparison.

V. Results Analysis

To evaluate the effectiveness of our proposed MGRN framework, we conducted comprehensive experiments on ten benchmark datasets. Table I presents the performance comparison between MGRN and various competitive baseline models. The results clearly demonstrate that pre-trained models significantly outperform non-pre-trained approaches, which can be attributed to their extensive learning corpus and powerful information extraction capabilities. When integrated with BERT-base and BERT-large, our MGRN framework achieves average accuracy improvements of 0.8% and 0.7%, respectively, across all datasets. These results substantiate the effectiveness of our multi-granularity reasoning approach for natural language inference tasks. Notably, MGRN outperforms RoBERTa-base by 1.5% and RoBERTa-large by 0.5%, indicating that our method can effectively model sentence relationships from both local and global perspectives, thereby capturing more fine-grained and complex semantic interactions. These improvements highlight the advantage of multi-level modeling in semantic mining, as it simultaneously captures both local and global differential information. Compared to previous state-of-the-art approaches, our method demonstrates highly competitive performance across all semantic similarity evaluation tasks, further validating the effectiveness of the proposed framework.

A. Ablation Study

To evaluate the individual contribution of each component in our MGRN framework, we conducted comprehensive ablation studies showed in Table II. Removing the interaction matrix reduces accuracy (85.1 $\rightarrow$ 84.7 on matched), showing its effectiveness in capturing fine-grained semantic relations. Replacing DenseNet with average pooling causes a clear drop (1.2% on matched), highlighting its importance for high-level feature extraction. Using only the last BERT layer degrades performance, confirming that multi-layer fusion better preserves semantic information. Overall, the ablation results show that MGRN benefits from the joint contribution of the interaction matrix, DenseNet, and multi-layer stacking.

B. Robustness Test Performance

To further evaluate the robustness of our proposed MGRN framework, we conduct adversarial and perturbation-based robustness tests on three widely studied datasets, including Quora Question Pairs (QQP), SNLI, and MNLI. Table III reports the performance of MGRN and several competitive

TABLE I
PERFORMANCE COMPARISON OF MGRN WITH OTHER METHODS ON TEN BENCHMARK DATASETS.

Model	Pre-train	MRPC	QQP	STS-B	MNLI-m/mm	QNLI	RTE	SNLI	Sci	SICK	Twi	Avg
BiMPM	✗	79.6	85.0	-	72.3/72.1	81.4	56.4	-	-	-	-	-
CAFE	✗	82.4	88.0	-	78.7/77.9	81.5	56.8	88.5	83.3	72.3	-	-
ESIM	✗	80.3	88.2	-	75.8/75.6	80.5	-	88.0	70.6	71.8	-	-
Transformer	✗	81.7	84.4	73.6	72.3/71.4	80.3	58.0	84.6	72.9	70.3	68.8	74.4
BiLSTM+ELMo+Attnt	✓	84.6	86.7	73.3	76.4/76.1	79.8	56.8	89.0	85.8	78.9	81.4	78.9
OpenAI GPT	✓	82.3	81.3	80.0	82.1/81.4	87.4	56.0	88.4	84.8	79.5	81.9	80.4
UERBERT	✓	88.3	90.5	85.1	84.2/83.5	90.6	67.1	90.8	92.2	87.8	86.2	86.0
SemBERT	✓	88.2	90.2	87.3	84.4/84.0	90.9	69.3	90.9	92.5	87.9	86.8	86.5
SyntaxBERT	✓	89.2	89.6	88.1	84.9/84.6	91.1	68.9	91.0	92.7	88.7	87.3	86.3
MGRN	✓	89.1	91.3	88.2	84.9/84.7	91.4	69.5	91.3	93.6	88.6	87.5	86.7
BERT-Base	✓	87.2	89.1	86.8	84.3/83.7	90.4	67.2	90.7	91.8	87.2	84.8	85.8
BERT-Base-MGRN	✓	**89.3**	**89.1**	**87.1**	**85.1/84.9**	**91.2**	**68.5**	**91.2**	**92.1**	**87.8**	**86.6**	**86.8**
BERT-Large	✓	88.9	89.3	87.6	86.8/86.3	92.7	70.1	91.0	94.4	91.1	91.5	88.0
BERT-Large-MGRN	✓	**89.9**	**90.2**	**88.1**	**86.8/86.7**	**93.0**	**72.0**	**91.1**	**94.3**	**91.2**	**92.4**	**88.8**
RoBERTa-Base	✓	89.3	89.6	87.4	86.3/86.2	92.2	73.6	90.8	92.3	87.9	85.9	87.6
RoBERTa-Base-MGRN	✓	**89.5**	**91.2**	**88.1**	**87.7/87.5**	**93.2**	**82.5**	**91.2**	**93.1**	**89.5**	**87.6**	**89.2**
RoBERTa-Large	✓	89.4	89.7	90.2	89.5/89.3	92.7	83.8	91.2	94.3	91.2	91.9	90.3
RoBERTa-Large-MGRN	✓	**90.2**	**91.2**	**90.6**	**90.2/90.1**	**94.2**	**84.1**	**91.2**	**94.2**	**91.4**	**92.3**	**90.7**

TABLE II
RESULTS OF ABLATION ON MULTINLI DATASET.

Ablation Experiments	Dev Accuracy	
	Matched	Mismatched
MGRN_{bert_base} (Full model)	85.1	84.9
- w/o multi-layer interaction	84.6	83.9
- w/o interaction matrix	84.7	84.1
- w/o DenseNet extraction	83.9	83.4

baseline models under different transformation settings. Overall, MGRN consistently achieves superior or highly competitive performance across most perturbation types, demonstrating its strong robustness and generalization capability. In the SwapAnt transformation, which introduces semantic contradictions by replacing words with their antonyms, MGRN significantly outperforms all baseline models on QQP. This result highlights the advantage of explicitly modeling fine-grained interaction patterns across multiple representation layers, enabling MGRN to better capture semantic polarity shifts that are difficult to identify using surface-level alignments alone. For the NumWord transformation, which requires precise numerical reasoning, all models experience performance degradation. However, MGRN outperforms BERT by a clear margin, indicating that multi-granularity interaction representations help preserve subtle semantic differences introduced by numerical changes. Under synonym-based perturbations such as SwapSyn, UERBERT benefits from explicitly injected lexical knowledge and achieves strong performance. Notably, MGRN attains comparable results without relying on external knowledge resources, suggesting that hierarchical semantic interactions encoded in intermediate transformer layers already provide rich contextual cues for semantic equivalence reasoning. In transformations that disrupt surface form or syntactic regularity, such as TwitterType and AddPunc, models that heavily depend on syntactic structures (e.g., SyntaxBERT) exhibit noticeable performance drops. In contrast, MGRN maintains stable performance, indicating that its reasoning process relies more on semantic interaction patterns rather than fragile surface or syntactic features.

C. Case Study

To qualitatively analyze how MGRN performs fine-grained reasoning, we present three representative examples in Table IV. These cases involve subtle lexical, gender-related, and numerical differences that often challenge conventional NLI models. In the first example, the non-pretrained ESIM model fails to recognize the semantic conflict caused by subtle word substitutions, resulting in incorrect predictions. While BERT benefits from contextualized representations and correctly predicts this case, it struggles in the third example, where numerical differences (e.g., “12” vs. “42”) require precise semantic discrimination. SyntaxBERT enhances text understanding by incorporating syntactic structures; however, in cases where sentence pairs share highly similar syntactic patterns, such as the second and third examples, syntactic cues alone are insufficient to distinguish semantic differences, leading to incorrect predictions. In contrast, MGRN consistently makes correct predictions across all cases. By explicitly modeling fine-grained interaction patterns across multiple transformer layers, MGRN is able to capture subtle semantic discrepancies introduced by lexical substitutions, gender changes, and numerical variations. This demonstrates the advantage of multi-granularity reasoning in accurately identifying semantic relations that cannot be resolved by surface-level or single-layer representations alone.

Model	Quora					SNLI				
	SA	NW	IA	Al	BT	AS	SA	TT	SN	SW
ESIM† [4]	-	-	-	-	-	64.00	84.22	78.32	53.76	65.38
BERT‡ [9]	48.58	56.96	86.32	**85.48**	83.42	79.66	94.84	83.56	50.45	76.42
ALBERT‡ [**?**]	51.08	55.24	81.87	78.94	82.37	45.17	96.37	81.62	57.66	74.93
UERBERT‡ [23]	48.57	54.86	84.72	80.88	82.71	73.24	94.78	85.36	57.54	80.81
SemBERT‡ [15]	50.92	53.15	85.19	82.04	82.40	76.81	95.31	84.60	56.28	77.86
SyntaxBERT‡ [**?**]	49.30	56.37	86.43	84.62	84.19	78.63	95.02	**86.91**	58.26	76.90
MGRN‡	**60.43**	**62.76**	**87.50**	85.48	**87.49**	**81.06**	**96.85**	85.14	**60.58**	**80.92**

Method	MNLI-m/mm					
	AS	SA	AP	TT	SN	SW
BERT‡ [9]	55.32/55.25	52.76/55.69	82.30/82.31	77.08/77.22	51.97/51.84	76.41/77.05
ALBERT‡ [**?**]	53.09/53.58	50.25/50.20	**83.98/83.68**	**77.98**/78.03	56.43/50.03	76.63/77.43
UERBERT‡ [23]	54.99/54.84	52.29/53.80	79.80/79.18	75.46/74.93	55.21/55.96	**82.23**/82.74
SemBERT‡ [15]	55.38/55.12	54.07/54.62	78.70/78.16	73.90/73.47	53.43/53.76	78.09/78.93
SyntaxBERT‡ [**?**]	54.92/54.63	53.54/54.73	77.01/76.71	70.38/70.13	57.11/51.95	78.57/79.31
MGRN‡	**60.14/59.25**	**60.89/61.37**	83.23/83.19	77.94/**78.10**	**60.12/59.83**	82.15/**82.97**

TABLE III

ROBUSTNESS EVALUATION RESULTS OF MGRN AND BASELINE MODELS UNDER VARIOUS ADVERSARIAL AND PERTURBATION SETTINGS. THE APPLIED DATA TRANSFORMATION METHODS INCLUDE SWAPANT (SA), NUMWORD (NW), ADDSENT (AS), INSERTADV (IA), APPENDIRR (AL), ADDPUNC (AP), BACKTRANS (BT), TWITTERTYPE (TT), SWAPNAMEDENT (SN), AND SWAPSYN-WORDNET (SW).

Case	ESIM	BERT	SyntaxBERT	MGRN
S1:How done you solve this aptitude question? S2:How does I solve aptitude questions on cube?	label:1	label:0	label:0	similarity:10.87% label:0
S1:How can I tell if this girl loves me? S2:How can I tell if this boy loves me?	label:1	label:1	label:1	similarity:12.06% label:0
S1:How many 12 digits number have the sum of 4? S2:How many 42 digits number have the sum of 4?	label:1	label:1	label:1	similarity:18.63% label:0

TABLE IV

THE EXAMPLE SENTENCE PAIRS OF OUR CASES. RED AND BLUE ARE DIFFERENCE PHRASES IN SENTENCE PAIR.

VI. CONCLUSION

In this paper, we proposed a novel Multi-Granularity Reasoning Network (MGRN) for natural language inference that leverages hierarchical semantic features through an interactive tensor structure. Our approach mimics the human cognitive process of language understanding, progressing from shallow to deep levels of comprehension to effectively uncover intricate semantic information behind linguistic expressions. Extensive experiments on multiple public benchmarks demonstrate that MGRN consistently outperforms strong baseline models across various NLI tasks.

REFERENCES

[1] M. Wang, N. A. Smith, and T. Mitamura, "What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 22–32, 2007.

[2] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A Large Annotated Corpus for Learning Natural Language Inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 632–642, 2015.

[3] K. He, X. Zhang, R. Shaoqing, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[4] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced LSTM for Natural Language Inference," *arXiv preprint arXiv:1609.06038*, 2016.

[5] R. Jia and P. Liang, "Adversarial Examples for Evaluating Reading Comprehension Systems," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2021–2031, 2017.

[6] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," *arXiv preprint arXiv:1705.02364*, 2017.

[7] Y. Gong, H. Luo, and J. Zhang, "Natural Language Inference over Interaction Space," *arXiv preprint arXiv:1709.04348*, 2017.

[8] J. Choi, K. M. Yoo, and S.-g. Lee, "Learning to Compose Task-Specific Tree Structures," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171–4186, 2019.

[10] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5753–5763, 2019.

[12] D. Liang, F. Zhang, Q. Zhang, and X.-J. Huang, "Asynchronous deep interaction network for natural language inference," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 2692–2700, 2019.

[13] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[15] Z. Zhang, Y. Wu, H. Zhao, L. Zuchao, S. Zhang, X. Zhou, and X. Zhou, "Semantics-Aware BERT for Language Understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 9628–9635, 2020.

[16] S. Peng, H. Cui, N. Xie, S. Li, J. Zhang, and X. Li, "Enhanced-RCNN: An Efficient Method for Learning Sentence Similarity," in *Proceedings of The Web Conference (WWW)*, pp. 2500–2506, 2020.

[17] W. Ma, Y. Cui, C. Si, T. Liu, S. Wang, and G. Hu, "CharBERT: Character-Aware Pre-Trained Language Model," in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 39–50, 2020.

[18] Z. Hu, B. Tan, and J. Luo, "Heterogeneous Graph Transformer for

Graph-to-Text Generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–12, 2020.

[19] Q. Guo, X. Xiao, T. Ma, et al., "Graph neural networks for natural language processing: A survey," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[20] A.-L. Collomb, A. Conesa-Bes, and L. Lautman, "Artificial Intelligence and Law: Risks, Research, and Limits," *Artificial Intelligence and Law*, vol. 28, 2020.

[21] F. Christopoulou, M. Miwa, and S. Ananiadou, "Connecting the dots: Document-level relation extraction with edge-oriented graphs," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[22] S. Xu, S. E, and Y. Xiang, "Enhanced Attentive Convolutional Neural Networks for Sentence Pair Modeling," *Expert Systems with Applications*, 2020.

[23] T. Xia, Y. Wang, Y. Tian, and Y. Chang, "Using Prior Knowledge to Guide BERT's Attention in Semantic Textual Matching Tasks," in *Proceedings of the Web Conference (WWW)*, pp. 2466–2475, 2021.

[24] T. Gui, X. Wang, Q. Zhang, et al., "TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 347–355, 2021.

[25] P. Liu, J. Fu, Y. Xiao, W. Yuan, J. Chang, and Z. Geng, "Explainaboard: An Explainable Leaderboard for NLP," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 280–289, 2021.

[26] Chang Dai, Hongyu Shan, Mingyang Song, and Di Liang. Hope: Hyperbolic rotary positional encoding for stable long-range dependency modeling in large language models. *arXiv preprint arXiv:2509.05218*, 2025.

[27] Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuan-Jing Huang. Cqg: A simple and effective controlled generation framework for multi-hop question generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6896–6906, 2022.

[28] Ziyuan Gao, Di Liang, Xianjie Wu, Philippe Morel, and Minlong Peng. Decorl: Decoupling reasoning chains via parallel sub-step generation and cascaded reinforcement for interpretable and scalable rlhf. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30789–30797, 2026.

[29] Tao Gui, Qi Zhang, Jingjing Gong, Minlong Peng, Di Liang, Keyu Ding, and Xuan-Jing Huang. Transferring from formal newswire domain with hypernet for twitter pos tagging. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2540–2549, 2018.

[30] Bo Li, Di Liang, and Zixin Zhang. Comateformer: Combined attention transformer for semantic sentence matching. *arXiv preprint arXiv:2412.07220*, 2024.

[31] Junchen Li, Chao Qi, Rongzheng Wang, Qizhi Chen, Liang Xu, Di Liang, Bob Simons, and Shuang Liang. When safety becomes a vulnerability: Exploiting llm alignment homogeneity for transferable blocking in rag. *arXiv preprint arXiv:2603.03919*, 2026.

[32] Liang Li, Qisheng Liao, Meiting Lai, Di Liang, and Shangsong Liang. Local and global: Text matching via syntax graph calibration. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11571–11575. IEEE, 2024.

[33] Di Liang, Fubao Zhang, Weidong Zhang, Qi Zhang, Jinlan Fu, Minlong Peng, Tao Gui, and Xuanjing Huang. Adaptive multi-attention network incorporating answer information for duplicate question detection. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 95–104, 2019.

[34] Peiyang Liu, Ziqiang Cui, Di Liang, and Wei Ye. Who stole your data? a method for detecting unauthorized rag theft. *arXiv preprint arXiv:2510.07728*, 2025.

[35] Xiaoyu Liu, Xiaoyu Guan, Di Liang, and Xianjie Wu. Dpi: Exploiting parameter heterogeneity for interference-free fine-tuning. *arXiv preprint arXiv:2601.17777*, 2026.

[36] Xiaoyu Liu, Di Liang, Hongyu Shan, Peiyang Liu, Yonghao Liu, Muling Wu, Yuntao Li, Xianjie Wu, Li Miao, Jiangrong Shen, et al. Structural reward model: Enhancing interpretability, efficiency, and scalability in reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 672–685, 2025.

[37] Yonghao Liu, Mengyu Li, Di Liang, Ximing Li, Fausto Giunchiglia, Lan Huang, Xiaoyue Feng, and Renchu Guan. Resolving word vagueness with scenario-guided adapter for natural language inference. *arXiv preprint arXiv:2405.12434*, 2024.

[38] Yonghao Liu, Di Liang, Fang Fang, Sirui Wang, Wei Wu, and Rui Jiang. Time-aware multiway adaptive fusion network for temporal knowledge graph question answering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[39] Yonghao Liu, Di Liang, Mengyu Li, Fausto Giunchiglia, Ximing Li, Sirui Wang, Wei Wu, Lan Huang, Xiaoyue Feng, and Renchu Guan. Local and global: Temporal question answering via information fusion. In *IJCAI*, pages 5141–5149, 2023.

[40] Ruotian Ma, Yiding Tan, Xin Zhou, Xuanting Chen, Di Liang, Sirui Wang, Wei Wu, Tao Gui, and Qi Zhang. Searching for optimal subword tokenization in cross-domain ner. *arXiv preprint arXiv:2206.03352*, 2022.

[41] Qi Qian, Muling Wu, Zisu Huang, Wenhao Liu, Changze Lv, Xiaohua Wang, Zhenghua Wang, Zhengkang Guo, Zhibo Xu, Lina Chen, et al. Adaptive curriculum strategies: Stabilizing reinforcement learning for large language models.

[42] Jian Song, Di Liang, Rumei Li, Yuntao Li, Sirui Wang, Minlong Peng, Wei Wu, and Yongxin Yu. Improving semantic matching through dependency-enhanced pre-trained model with adaptive fusion. *arXiv preprint arXiv:2210.08471*, 2022.

[43] Rongzheng Wang, Yihong Huang, Muquan Li, Jiakai Li, Di Liang, Bob Simons, Pei Ke, Shuang Liang, and Ke Qin. Rethinking llm-driven heuristic design: Generating efficient and specialized solvers via dynamics-aware optimization. *arXiv preprint arXiv:2601.20868*, 2026.

[44] Sirui Wang, Di Liang, Jian Song, Yuntao Li, and Wei Wu. Dabert: Dual attention enhanced bert for semantic matching. *arXiv preprint arXiv:2210.03454*, 2022.

[45] Yao Wang, Di Liang, and Minlong Peng. Not all parameters are created equal: Smart isolation boosts fine-tuning performance. *arXiv preprint arXiv:2508.21741*, 2025.

[46] Muling Wu, Qi Qian, Wenhao Liu, Xiaohua Wang, Zisu Huang, Di Liang, LI Miao, Shihan Dou, Changze Lv, Zhenghua Wang, et al. Progressive mastery: Customized curriculum learning with guided prompting for mathematical reasoning. *arXiv preprint arXiv:2506.04065*, 2025.

[47] Xianjie Wu, Di Liang, Jian Yang, Xianfu Cheng, LinZheng Chai, Tongliang Li, Liqun Yang, and Zhoujun Li. Breaking size barrier: Enhancing reasoning for large-size table question answering. In *International Conference on Database Systems for Advanced Applications*, pages 241–256. Springer, 2025.

[48] Xianjie Wu, Xiaohang Xu, Tingyu Jiang, Jian Yang, Di Liang, Xianfu Cheng, Zhenhe Wu, Linzheng Chai, Wei Zhang, Jiaheng Liu, et al. Mmtablebench: A multi-level multimodal benchmark for reasoning and layout complexity in table qa. In *Proceedings of the ACM Web Conference 2026*, pages 3881–3892, 2026.

[49] Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25497–25506, 2025.

[50] Xianjie Wu, Jian Yang, Tongliang Li, Shiwei Zhang, Yiyang Du, LinZheng Chai, Di Liang, and Zhoujun Li. Unleashing potential of evidence in knowledge-intensive dialogue generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[51] Chao Xue, Di Liang, Pengfei Wang, and Jing Zhang. Question calibration and multi-hop modeling for temporal question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19332–19340, 2024.

[52] Chao Xue, Di Liang, Sirui Wang, Jing Zhang, and Wei Wu. Dual path modeling for semantic matching by perceiving subtle conflicts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[53] Rui Zheng, Rong Bao, Yuhao Zhou, Di Liang, Sirui Wang, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Robust lottery tickets for pre-trained language models. *arXiv preprint arXiv:2211.03013*, 2022.