Human-Centric Intelligent Systems (2024) 4:127–135
https://doi.org/10.1007/s44230-023-00054-y

RESEARCH ARTICLE

The Linguistic Feature Relation Analysis of Premise and Hypothesis for Interpreting Nature Language Inference

Xinyu Chen[1] · Lin Li[1] · Mengjing Zhang[1] · Rui Zhang[1]

Received: 28 August 2023 / Accepted: 22 November 2023 / Published online: 22 March 2024

Abstract
Natural language inference (NLI) is a fundamental task of natural language processing (NLP). Most recent NLI research has focused on explaining the model's decisions in generating causal explanations (i.e., why did a premise/hypothesis pair as input lead to their inference relation as output?). As layer-based language models can learn language structure information, this paper conducts a sample-by-sample analysis of the linguistic feature relation between premise and hypothesis that is expected to guide NLI modeling and interpretation better. Our empirical study verifies that the linguistic feature relation of premise/hypothesis pairs can be seen in NLI inference models, which can be used to interpret inference samples. Meanwhile, experimental results show that these linguistic features relation interpretation can help the NLI model achieve comparable inference accuracy compared with state-of-the-art methods.

Keywords Nature language inference · Interpretability · Linguistic feature

1 Introduction

As a widely studied task in natural language processing (NLP), natural language inference (NLI) determines whether a hypothesis sentence can be inferred from a premise sentence [1, 2]. NLI models predict relationships between premise and hypothesis pairs, deciding whether a hypothesis is contained by the premise (i.e., neutral, contradiction, entailment).

Deep learning (DL) models have become increasingly popular in fields such as finance and medicine due to their ability to learn large amounts of information from data and achieve high accuracy. Explainable artificial intelligence (XAI) which provides interpretability for deep learning neural networks has received increasing attention since black-box-based representations in deep learning will reduce the trust of end users and hinder its further development and application [3, 4]. It is also found in the development of XAI that the addition of explainable methods may weaken the accuracy of deep learning models, and the main question gradually turns to whether we can get a highly accurate interpretable model comparable to the accuracy provided by deep learning models [5].

Recently, much of the interpretability work in NLI attempts to provide users with causal explanations, such as outputting specific labels through automatic generation [6]. Counterfactual examples are used to provide contrastive explanations for labels [7]. Some causal explanations are in the form of visual analysis [8]. These works supply users with well-labeled clues, focusing on explaining why a premise/hypothesis pair is judged to be a specific inference relation [9]. Different from causal explanations, we aim to build a model that provides self-explanation for the process of judgment.

Our analysis found that when humans do interpreting tasks, they make judgments based on the linguistic feature relation between sentence pairs (see Fig. 1). Such linguistic

Xinyu Chen, Lin Li, Mengjing Zhang, and Rui Zhang have contributed equally to this work.

✉ Xinyu Chen
268595@whut.edu.com

Lin Li
cathylilin@whut.edu.cn

Mengjing Zhang
zhangmengijng@whut.edu.cn

Rui Zhang
zhangrui@whut.edu.cn

[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Luoshi Street, Wuhan 430070, Hubei, China

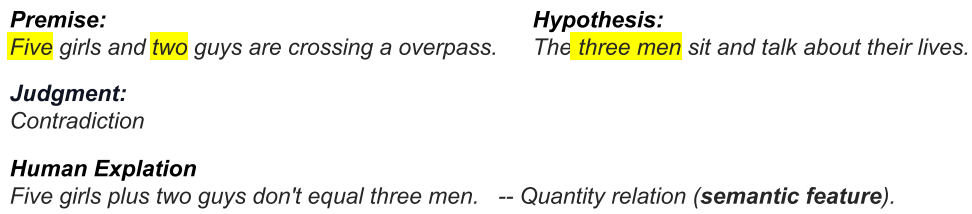

Fig. 1 An sample of using linguistic structures to explain the basis for a judgment

feature relation can be used to understand and infer the judgment of a premise/hypothesis pair. We expect to incorporate this human interpretation to inspire the model to better learn linguistic feature relations.

Existing research shows that layer-based language models can capture language structures in layers [10]. Thus, this paper conducts a sample-by-sample study to explore the linguistic feature relation of sentence pairs, which is expected to guide NLI model design better. Based on this, a multi-layer connection mechanism [11] is added to the model in order to learn how humans use linguistic feature relations. This multi-layer connection mechanism can learn the weights of each layer to adapt to the inference samples of various linguistic feature relations and explain which linguistic feature relation our model focuses on to make predictions. Through our empirical research, we confirm that:

- When NLI uses phrase feature relations (e.g., date, noun correspondence) to explain, it uses more of the language structure learned by lower layers.
- When NLI uses syntactic feature relations (e.g., attributive clause, adverbial) to explain, it uses more of the language structure learned by middle layer.
- When NLI uses semantic feature relations (e.g., attributive clause, adverbial) to explain, it uses more of the language structure learned by higher layer.

Additionally, our experimental results show that this interpretation based on linguistic feature relation can achieve comparable accuracy to state-of-the-art methods for XAI models.

2 Related Work

In recent years, as deep learning increasingly affects the various aspects of society and life, making the interpretability of deep learning models crucial [12].

Table 1 The number distribution of each label in the SNLI dataset

Dataset	*Entailment*	*Neutral*	*Contradiction*
SNLI	3368	3219	9824
Hard SNLI	1058	1068	1135
Easy SNLI	2310	2151	2102

Linguistic Structures. Some experiments focus on what types of information the model can capture because of the current popularity of multi-layer language models. Jawahar et al. [10] study some probing tasks including span representation [13] and sentence embedding [14] detection to explore the representation of linguistic structures learned by BERT.

Interpretability. How to provide users with clues about the model's prediction results is the focus of recent interpretability work on NLI tasks [7, 8, 15]. For example, some researchers focus on interpreting the predictions of neural models in a model-agnostic manner [15]. Following these researches, Chen et al. [7] use counterfactual examples to add an explanation of "why A and not B" to the model prediction results. Kalouli [8] provides interpretability in the form of visual analytics. On the other hand, adding attention mechanisms is also one of the ways to provide interpretability for deep learning models. As in works done by Park et al. [16] and Vig et al. [17] which is using this mechanism to help build deep learning models with self-generated explanations.

Since the detection tasks in BERT [10] are all based on a single sentence, it is unclear whether layer-based language models can capture the linguistic feature relation between sentence pairs when dealing with NLI tasks.

3 Dataset

In our study, the SNLI dataset is used to compare with state-of-the-art NLI methods [18]. To better evaluate the robustness of the model, this paper increases the use of Hard SNLI and Easy SNLI datasets in ablation experiments. It introduces these three datasets separately in this section.

SNLI Dataset. Stanford proposed Natural Language Inference (SNLI)[1] dataset in 2015 as the first large-scale artificially annotated dataset for natural language inference tasks. It contains 570k human-annotated sentence pairs, 550k for training pairs, 10k for test pairs, and 10k for development pairs.

Easy/Hard SNLI Dataset. Gururangan et al. [19] found that since the construction and labeling of the SNLI dataset are done manually, it may contain a lot of information that

[1] https://nlp.stanford.edu/projects/snli/

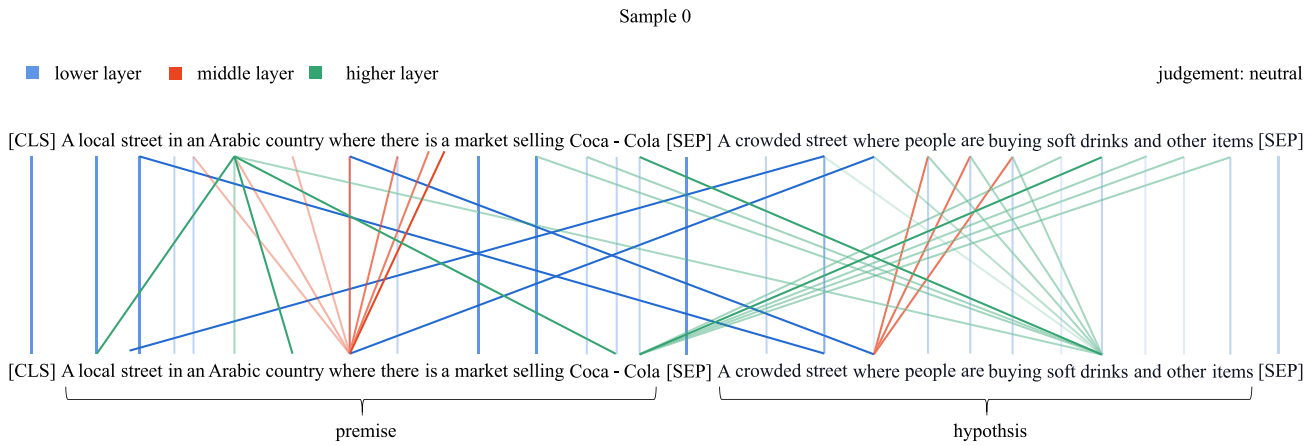

Fig. 2 The attention view of the same sample about different linguistic feature relations learned by the layers (where the color is used to represent the attention of different layers, and the color depth represents the weight of the attention)

guides the model to the correct result. Experiments show that evaluating the model with only hypothesis sentences from the SNLI dataset can achieve 67% accuracy. These rules make each label in the SNLI dataset have its own characteristics. Therefore, Gururangan et al. [19] split two test sets, Hard SNLI and Easy SNLI, based on the SNLI dataset. The Hard SNLI dataset is formed by removing all sentence pairs containing regularities. The details of SNLI dataset are shown in Table 1.

4 Linguistic Feature Relation Analysis

Since an inference sample in NLI contains two sentences that rely on various linguistic feature relations for predicting judgments, this section conducts a validation analysis of samples in the SNLI dataset to explore the relationship between samples, language structures, and model layers.

4.1 Linguistic Feature Relation Between Premise and Hypothesis

Jawahar et al. [10] stated the ability of the layer-based models to capture language structures through detection experiments: lower layer can capture phrase features; middle layer can capture syntactic features; higher layer can capture semantic features. In order to better incorporate human explanation into NLI models, we propose to explore how the linguistic feature relations between sentence pairs are captured by model layers. Following the latest baseline experiments, Roberta [20] was used to explore feature relationships.

50 samples are selected in the SNLI dataset so that those samples are uniform in label (entailment, neutral, contradiction) distribution. And then we analyze the attention heads of all encoding layers for each sample. By comparing the attention mechanism weights of different layers, it is verified that different layers can capture different linguistic structures.

In Fig. 2, we first pass the attention of the same sentence in different layers and find that the semantic feature relations will be captured in the form of attention view in the sentence pair. And the phrase/syntactic/semantic feature relations are learned at lower, middle, and higher layers respectively.

It can be seen from the figure that:

- The model learns phrase feature relations in premise and hypothesis sentences in lower layer.
- In middle layer, the model analyzes the “where” clause, aligns the components of the clause and focuses on the syntactic feature relations.
- In higher layer, the model focuses on parsing semantic relations between words such as “coca-cola” and “soft drinks”.

Therefore, it can be found that the layer-based NLI model learns the linguistic feature relation of premises and hypothesis with the help of the attention mechanism. The learning of the linguistic feature relation between the sentence pairs is achieved by adjusting the weights of words between the premise sentences and the hypothesis sentences in each layer.

4.2 Interpreting with Linguistic Feature Relation

In Fig. 3 we analyze examples using different linguistic feature relations (phrasal/ syntactic/ semantic) for inference.

Sample 1 The clues of judging entailment are dependent mainly on phrase feature relation (the alignment of the

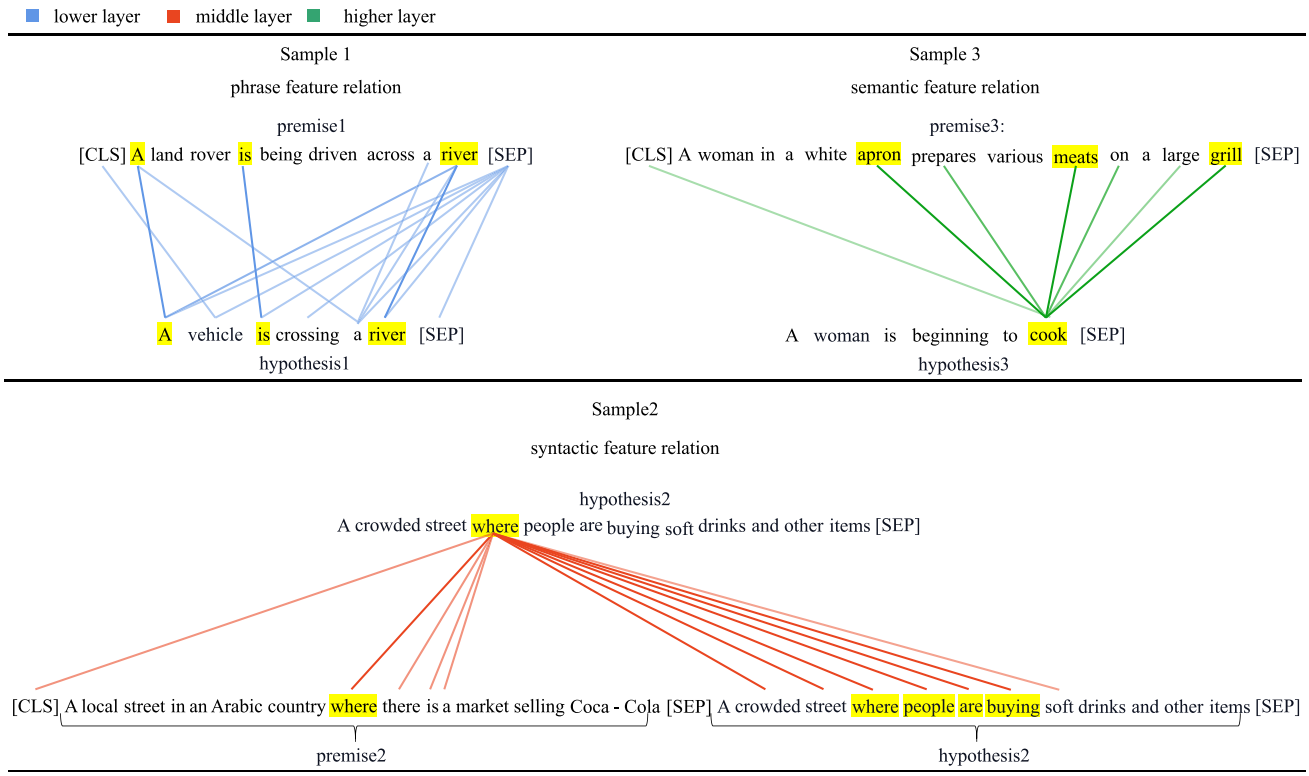

Fig. 3 Attention view of lower/middle/higher layer samples in NLI model. The depth of the blue line indicates the size of the attention

numeral "a", the alignment of the noun "river" and the tense-related be verb "is"). Such phrase feature relation is captured by lower layers as shown in the upper left part of Fig. 3. It is clear that the alignment relationship of words between sentence pairs is in lower layers.

Sample 2 In the lower part of Fig. 3, the premise and hypothesis sentences are both attributive clauses guided by "where". In middle layers, the "where" of the hypothesis sentence forms a corresponding relationship with the "where" in the premise sentence and the phrases involving attributive clauses ("a crowded street" and "people are buying") in the hypothesis sentence.

Sample 3 In the upper right part of Fig. 3, it is shown that higher layers of the NLI model are more suitable for processing rich semantic feature relations. As in this sample, the judgment of "cook" in the hypothesis sentence relies on the phrases that do not directly refer to cooking but are involved in cooking (e.g., "apron", "various meats", "grill") in the premise sentence. This semantic correspondence enables alignment at higher layers.

Through the analysis of samples, it can be summarized as follow:

- We verify that the linguistic feature relation in premise/hypothesis pairs can be learned in NLI model, such as phrase feature relations captured by lower layers, syntactic feature relations captured by middle layer and semantic feature relations captured by higher layers.
- Since the linguistic feature relation between sentence pairs plays a different role(capturing phrase/syntactic/semantic features) in interpreting, lower/middle/higher layers of an NLI model will have different contributions to the judgment of premise/hypothesis pairs.

5 Experiment

Inspired by samples similar to those described in Sect. 3, it is evident that human interpretation can be achieved in the model through the distinction of attention to linguistic structural relations in the hierarchy. Therefore, we hope to add a multi-layer connection attention mechanism to let the model learn how to adjust the impact of language structure relationships in different samples.

As shown in Fig. 3 in Sect. 4, human explanation can be achieved in the NLI model through the distinction of attention to the linguistic feature relation in layers. Since the attention mechanism can learn the weights of the

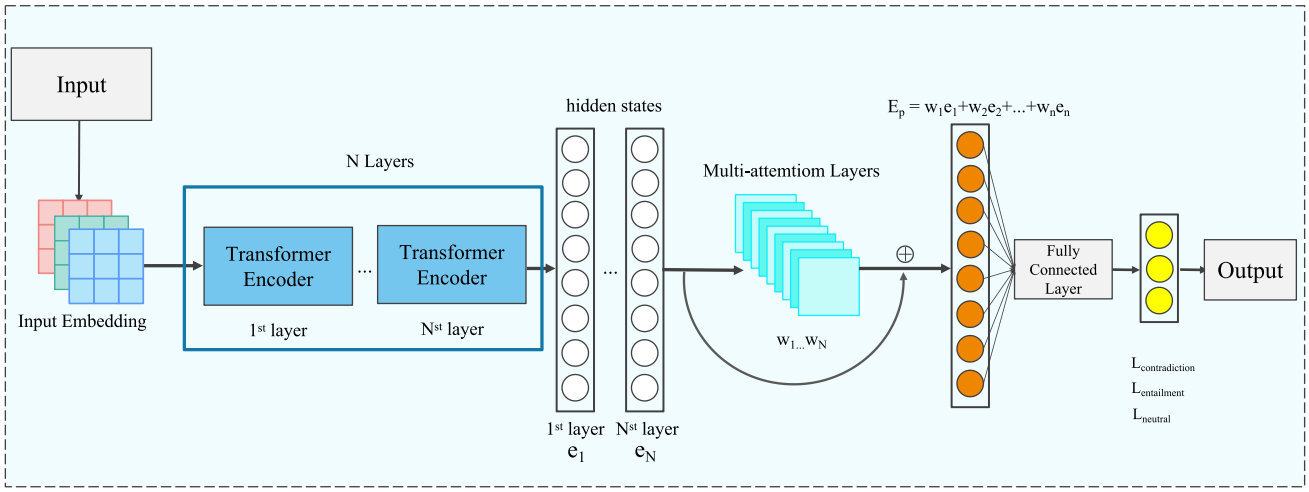

Fig. 4 Multi-layer connected NLI

input vector features of each layer [21]. Thus, we hope to add a multi-layer connection attention mechanism to allow the NLI model to adjust which linguistic feature relation should be focused on when interpreting different samples.

5.1 Multi-layer Connection Based NLI

In layer-based models (e.g., BERT, RoBERTa, XLNet, GPT), each layer is sequentially connected [20, 22–24]. In Fig. 4, our multi-layer connection based NLI model is divided into four parts as described below.

In the input layer, S_{Pi} (a premise pair) and S_{Hi} (a hypothesis pair) are put into the NLI model. And in the embedding layer, it is combined token embedding, segment embedding, and position embedding. Then the embedded input sequence is fed into the multi-attention layer.

In the multi-attention layer, this paper uses an attention mechanism to capture the attention relationship between different layers. Its core is to learn weights for the features of each layer via a linear connection. According to the weights learned by each layer, the coding representation of each layer of the hierarchical structure model is multiplied by the weights and then linearly connected to obtain the final embedding. Inference labels are predicted with softmax.

5.2 Interpretability Analysis

The attention mechanism score (w_i) of the i-th layer is got by dot product with K_i. Let $e = (e_0, e_1 \cdots e_{n-1})$ denote the feature vector output by the encoded hidden layer. In Eq. 1, the feature vector E of each layer with the attention mechanism weight W is multiplied, and the linear connection code with weights of each layer is calculated and represented as E_p.

$$E_p = w_0 \cdot e_0 + w_1 \cdot e_1 + \cdots + w_{n-1} \cdot e_{n-1} \quad (1)$$

In the output layer, a linear full connection after E_p is adopted in the prediction layer and softmax normalization is used to predict the judgments (entailment/neural/contradiction). In this way, the inference model with a multi-layer attention mechanism can realize layer-based interpretation by explaining which layer is the most concerned when making judgments for inference samples.

This paper builds upon the RoBERTa model by adding a multi-layer attention mechanism. Following Jawahar et al. [10], we define layers 1–4 as lower layers, layers 4–8 as middle layers, and layers 8–12 as higher layers.

The process of selecting samples is shown in Fig. 5 and Eq. 1. In Eq. 1, S_{pi} represents the premise sentence,S_{hi} represents the hypotheses sentence, $label_i$ represents the real label, $label_{pi}$ represents the predicted label.

First, according to the correctly predicted samples, find the layer that each sample is most dependent on (with the largest weight) during prediction, and divide the samples according to the layer. Divide the samples into three types of samples with the largest weight layer in the lower, middle, and higher layers. We randomly select samples from each type of sample and make a sample-by-sample analysis. In Table 2 we randomly choose some samples from these three inference labels of data for detailed analysis.

Sample 1: The judgment of this sentence pair is a contradiction. The layer with the largest weight in the model is from lower layers. The main basis of linguistic feature relation is that the instrument mentioned in the premise is "guitar" instead of "banjo", and the location in the premise

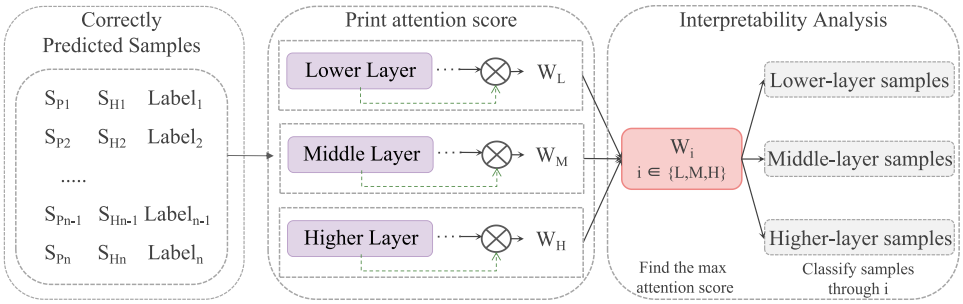

Fig. 5 The process of selecting inference samples

Table 2 Inference samples for analysis

No	*Premise*	*Hypothesis*	*Judgment*
1	A man playing an electric guitar on stage	A man playing banjo on the floor	Contradiction
2	A school girl is jumping over a low hurdle.	A girl is playing a sport in which she is jumping over a hurdle	Neutral
3	A woman wearing a tank top and black pants is laying on the ground while looking at a book	A woman is deciding whether or not to read a book	Neutral
4	A person rolls down a hill riding a wagon as another watches	A child in a wagon rolls down a hill	Entailment
5	A girl playing a violin along with a group of people. watches	A group of people are playing in a symphony	Neutral

is "stage" instead of "floor". The judgment is based on the unequal relationship between the above four nouns. The maximum weight of the sample is in lower layer, and the samples are classified in the S_{lower} set. This shows that our model uses the phrase feature relation to interpreting by paying more attention to lower layer.

Sample 2: The judgment of the two sentences is neutral. The middle layer gets the largest weight layer during inference. The premise is a simple sentence, and the hypothesis is an attributive clause guided by "in which" as a positional adverbial. The same phenomenon can be found in Sample 3, where the premise is a "while" leading clause. The largest weights in middle layers indicate that our model is consistent with the linguistic feature relation which is the human explanation focus on when interpreting.

Sample 4: When judging "a child" in the hypothesis needs to find whether the corresponding subject is "a person" or "another" in the premise. As "A person" can be "a child", it is judged to be entailment. When inferring, such layer with the largest weights is in higher layers. This verification shows that when it comes to complex semantic situations (e.g., multiple personal pronouns), it needs higher layers to perform semantic understanding to complete the inference.

Sample 5: The judgment is neutral. The layer with the largest weight is in the higher layers. It can be seen that it mainly depends on the correspondence of the subject's number. The subjects between the premise sentence ("a girl" and "group of people") and the hypothesis sentence ("a group of people") are the same. When humans explain the quantity relationship between multiple subjects, they need to understand the semantic feature relation between the sentence pair. Our model achieves the correct prediction of such sentence pairs by increasing the importance of semantic feature relations captured in higher layers.

Through this analysis, some conclusions are 1)through the layer where the maximum weight of each sample is located, it can determine which linguistic feature relation is more concerned in the judgment of an inference sample. 2)The weights generated by the multi-layer attention mechanism can explain the process of model interpreting in the way of linguistic feature relation; Our experiments show that the linguistic feature relation between premise and hypothesis can be learned through a multi-layer connected model in a human explanation way, and the model can achieve self-interpretation.

Table 3 Experimental results in terms of accuracy

Interpretability	*Model*	*SNLI Accuracy(%)*
Without Interpretation	DR-BiLSTM[26]	89.3
	DMAN[27]	89.6
	SemBERT[28]	91.9
	CA-MTL[29]	92.1
With Interpretation	*Ours*	92.2
Without Multi-layer Connection	Ours(Roberta)	90.9

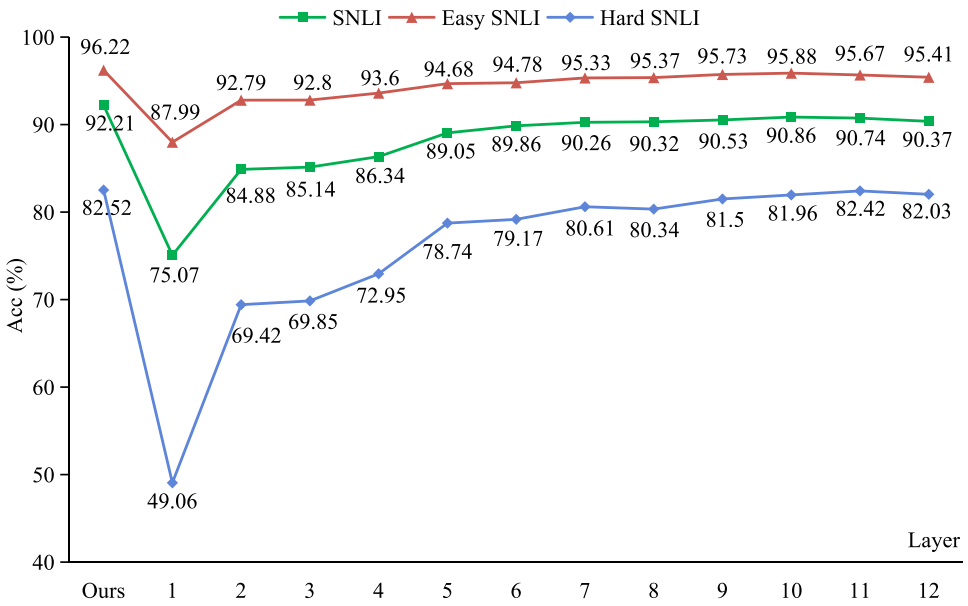

Fig. 6 The accuracy(%) results of each layer for prediction

5.3 Accuracy Results

For each baseline, we report the accuracy results in cited author papers. In our model, we set the learning rate to 1e-5 and the maximum sentence length to 128 according to the statistics of the sample length of the dataset, which is consistent across the three datasets (the SNLI/Easy SNLI/Hard SNLI dataset). We still use Adam Optimizer [25] to optimize our model parameters by computing different adaptive learning rates, setting the epoch to 6, and the model is saved every 10,000 steps. All experiments were performed using version 1.3.1 of Pytorch on a commodity machine equipped with 2 GTX 3090 and a total of 48GB memory.

Table 3 shows three classical approaches without the interpretation function for NLI tasks on the SNLI dataset. The experimental results are from the SNLI published results form[2] including:

- DR-BiLSTM [26], a model that improves the performance of NLI tasks through soft attention between premises and hypotheses.
- DMAN [27], a model based on an attention network to transfer knowledge through supervised tasks to solve NLI tasks.
- SemBERT [28], a multi-layer neural network model based on Transformer.
- CA-MTL [29], a model incorporating a new conditional attention mechanism and task conditioning module.

[2] https://nlp.stanford.edu/projects/snli/

By adding attention, the contribution of each layer to the final result can be explored, and the performance of the NLI task can be improved by training the model to fuse the information captured by all encoding layers. The accuracy results show that our model with a multi-layer attention mechanism can improve interpretability via learned linguistic structure, and has comparable accuracy performance compared with other methods.

5.4 Ablation Study

In order to analyze and compare the contributions of different network layers in our multi-layer neural network based NLI model, this paper designs the following ablation experiments on the SNLI dataset.

First, by removing Eq. 1 added in our model, it is verified that the multi-layer attention has an impact on the final prediction result. The experimental results show that the prediction accuracy of the model drops by 1.27% after removing the multi-layer attention mechanism. This shows that the model integrated with a human explanation way not only makes the model interpretable but also improves the accuracy of the model by learning the linguistic feature relation.

Second, each layer is used to be the input of the prediction layer for probing their prediction performance. As seen from Fig. 6, the prediction performance is poor with only 75% accuracy at lower layers. As the neural network layers become deep, from the first layer to the seventh layer, the prediction accuracy of the model increases rapidly. The performance difference between the two layers is about 1%. When the number of network layers reaches the seventh layer, the accuracy improvement of model prediction in the middle layer slows down.

It can be seen that the accuracy trends of different layer connection schemes on the SNLI, Hard SNLI, and Easy SNLI datasets are the same. In other words, as the number of layers increases, the accuracy of the model maintains a consistent upward trend. Since the difficulty of the test set in the Hard SNLI dataset is greater than that of the Easy SNLI dataset, it is more obvious in the upward trend of accuracy, highlighting the differences in the information captured by each encoding layer. This shows that the fusion of differential sentence encoding can capture richer and more extensive semantic information, which proves the effectiveness of our study of multi-layer language information for natural language reasoning tasks.

In summary, the ablation study shows that the performance and impact of each layer of the model on the prediction results are different. As the number of layers of the layer-based language model increases, the model may lose the phrase features and syntactic features captured by the lower and middle layers, resulting in the incorrect prediction of examples that rely more on these two features. The experimental results show that our NLI model with the multi-layer connection mechanism introduced in this paper can alleviate the problem that the information captured by lower layers is lost as the number of network layers goes larger. In addition, it also shows that the multi-layer attention-connected NLI model captures different information and contributes to the results differently.

6 Conclusion and Future Work

This paper focuses on interpreting NLI tasks with linguistic feature relation by adding a multi-layer connection mechanism. Our empirical study shows that linguistic feature relations can help NLI models interpret the prediction process in a way that humans explain. Conducted on the SNLI dataset, experimental results show that can multi-layer attention-based NLI model can make greater interpretability with 92.2% accuracy. In future work, we will increase the number of samples to further validate the linguistic structure in sentence pairs. Furthermore, for the linguistic feature relation captured by the lower, middle, and higher layers, we will try to make predictions via layer-wise prompt-based training.

Acknowledgements This work is partially supported by NSFC, China (No.62276196).

Author Contributions XC designed the work; LL interpreted the data; MZ created of new software used in the work; RZ drafted the work or substantively revised it.

Funding This work is partially supported by NSFC, China (No.62276196).

Availability of Data and Materials The datasets analysed during the current study are available in the SNLI repository, https://nlp.stanford.edu/projects/snli/.

Declarations

Conflict of Interest The authors declare that they have no competing interests. Author Lin Li is a member of the Editorial Board of Journal Human-Centric Intelligent Systems. The paper was handled by another Editor and has undergone a rigorous peer review process. Author Lin Li was not involved in the journal's peer review of, or decisions related to, this manuscript.

References

1. MacCartney B. Natural Language Inference. Stanford University; 2009.
2. Peters ME, Neumann M, Zettlemoyer L, Yih W. Dissecting contextual word embeddings: Architecture and representation. In: Proceedings of the 2018 Conference on EMNLP. ACL; 2018. p. 1499–1509.
3. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM International Conference in August 13–17, 2016. ACM; 2016. p. 1135–1144.
4. Pasquale F. The black box society: the secret algorithms that control money and information. Harvard University Press; 2015.
5. Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: an analytical review. WIREs Data Min Knowl Discov. 2021. https://doi.org/10.1002/widm.1424.
6. Kumar S, Talukdar PP. NILE: Natural language inference with faithful natural language explanations. In: Proceedings of the 58th Annual Meeting of ACL 2020. ACL; 2020. p. 8730–8742.
7. Chen Q, Ji F, Zeng X, Li F, Zhang J, Chen H, Zhang Y. KACE: generating knowledge aware contrastive explanations for natural language inference. In: Proceedings of the 59th Conference of ACL/IJCNLP 2021, (Volume 1: Long Papers). ACL; 2021. p. 2516–2527.
8. Kalouli A, Sevastjanova R, Paiva V, Crouch RS, El-Assady M. Xplainli: Explainable natural language inference through visual analytics. In: Proceedings of the 28th Conference of COLING 2020: System Demonstrations. ICCL; 2020. p. 48–52.
9. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter MA, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: 5th Conference of IEEE, DSAA 2018. IEEE; 2018. p. 80–89.
10. Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? In: Proceedings of the 57th Conference ACL 2019, Volume 1: Long Papers. ACL; 2019. p. 3651–3657.

11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. 2017; p. 5998–6008.
12. Socher R, Manning CD. Deep learning for NLP (without magic). In: Human Language Technologies: Conference of the North American Chapter of the ACL, Proceedings. ACL; 2013. p. 1–3.
13. Maaten L, Hinton G. Visualizing data using t-sne. J Mach Learn Res. 2008;9(11):2579–605.
14. Adi Y, Kermany E, Belinkov Y, Lavi O, Goldberg Y. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net 2017.
15. Cheng F, Ming Y, Qu H. DECE: decision explorer with counterfactual explanations for machine learning models. IEEE Trans Vis Comput Graph. 2021;27(2):1438–47.
16. Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M. Multimodal explanations: Justifying decisions and pointing to the evidence. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. IEEE Computer Society; 2018. p. 8779–8788.
17. Vig J. A multiscale visualization of attention in the transformer model. In: Costa-jussà MR, Alfonseca E (eds.) Proceedings of the 57th Conference of ACL 2019, Volume 3: System Demonstrations. ACL; 2019. p. 37–42.
18. Kim S, Kang I, Kwak N. Semantic sentence matching with densely-connected recurrent and co-attentive information. In: The Thirty-Third AAAI Conference on Artificial Intelligence. AAAI Press; 2019. p. 6586–6593.
19. Gururangan S, Swayamdipta S, Levy O, Schwartz R, Bowman SR, Smith NA. Annotation artifacts in natural language inference data. In: Proceedings of the 2018 Conference of NAACL-HLT, New Orleans, 2018. ACL; 2018. p. 107–112.
20. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: a robustly optimized BERT pretraining approach, 2019. CoRR arXiv:1907.11692.
21. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR, 2015.
22. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of NAACL-HLT 2019, Volume 1 (Long and Short Papers). ACL; 2019. p. 4171–4186.
23. Yang Z, Dai Z, Yang Y, Carbonell JG, Salakhutdinov R, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in NeurIPS 32: Annual Conference on NeurIPS 2019. 2019; p. 5754–5764
24. Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training, 2018.
25. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: 3rd ICLR 2015, Conference Track Proceedings, 2015.
26. Ghaeini R, Hasan SA, Datla VV, Liu J, Lee K, Qadir A, Ling Y, Prakash A, Fern XZ, Farri O. Dr-bilstm: Dependent reading bidirectional LSTM for natural language inference. In: Proceedings of the 2018 Conference of NAACL-HLT 2018, Volume 1 (Long Papers). ACL; 2018. p. 1460–1469.
27. Pan B, Yang Y, Zhao Z, Zhuang Y, Cai D, He X. Discourse marker augmented network with reinforcement learning for natural language inference. In: Proceedings of the 56th Annual Meeting of ACL 2018, Volume 1: Long Papers. ACL; 2018. p. 989–999.
28. Zhang Z, Wu Y, Zhao H, Li Z, Zhang S, Zhou X, Zhou X. Semantics-aware BERT for language understanding. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press; 2020. p. 9628–9635.
29. Pilault J, Elhattami A, Pal CJ. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters & less data. In: 9th ICLR 2021, Virtual Event, Austria, May 3–7, 2021. OpenReview.net, 2021.