BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 11/14 pp694–711
DOI: 10.26599/BDMA.2024.9020096
Volume 8, Number 3, June 2025

Disentangling Reasoning Factors for Natural Language Inference

Xixi Zhou†, Limin Zeng†, Ziping Zhao, Jiajun Bu, Wenjie Liang*, and Haishuai Wang*

Abstract: Natural Language Inference (NLI) seeks to deduce the relations of two texts: a premise and a hypothesis. These two texts may share similar or different basic contexts, while three distinct reasoning factors emerge in the inference from premise to hypothesis: entailment, neutrality, and contradiction. However, the entanglement of the reasoning factor with the basic context in the learned representation space often complicates the task of NLI models, hindering accurate classification and determination based on the reasoning factors. In this study, drawing inspiration from the successful application of disentangled variational autoencoders in other areas, we separate and extract the reasoning factor from the basic context of NLI data through latent variational inference. Meanwhile, we employ mutual information estimation when optimizing Variational AutoEncoders (VAE)-disentangled reasoning factors further. Leveraging disentanglement optimization in NLI, our proposed a Directed NLI (DNLI) model demonstrates excellent performance compared to state-of-the-art baseline models in experiments on three widely used datasets: Stanford Natural Language Inference (SNLI), Multi-genre Natural Language Inference (MNLI), and Adversarial Natural Language Inference (ANLI). It particularly achieves the best average validation scores, showing significant improvements over the second-best models. Notably, our approach effectively addresses the interpretability challenges commonly encountered in NLI methods.

Key words: Natural Language Processing (NLP); Natural Language Inference (NLI); disentangled representation learning; variational autoencoder

1 Introduction

Interpretable inference is a vital problem in Natural Language Processing (NLP) research for general Artificial Intelligence (AI). As its related fundamental task[1], Natural Language Inference (NLI) aims to determine the relations of two text segments: the premise and the hypothesis. Given a premise text, the hypothesis is of entailment if the NLI model can infer it from the premise, contradictory if the NLI model cannot infer it, and neutral if the NLI model cannot determine whether it is related to the premise text. Table 1 presents some basic examples of the NLI process.

Recently, various approaches have been developed to deduce the reasoning relation of the premise and the hypothesis. Most NLI models are based on the classical Siamese text-matching architecture[2] and the more advanced pair-wise interaction architecture, named the Bi-encoder and Cross-encoder, as demonstrated in Fig. 1.

• Xixi Zhou and Wenjie Liang are with Department of Radiology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China. E-mail: xixi.zxx@zju.edu.cn; baduen@zju.edu.cn.
• Limin Zeng, Jiajun Bu, and Haishuai Wang are with Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310027, China. E-mail: limin.zeng@zju.edu.cn; bjj@zju.edu.cn; haishuai.wang@zju.edu.cn.
• Ziping Zhao is with College of Computer Science, Tianjin Normal University, Tianjin 300387, China. E-mail: zhaoziping@tjnu.edu.cn.
† Xixi Zhou and Limin Zeng contribute equally to this work.
* To whom correspondence should be addressed.
Manuscript received: 2024-05-03; revised: 2024-11-20; accepted: 2024-12-03

Table 1 NLI examples. The reasoning factors (marked by color) determine the classification results of NLI relationships.

Premise (P) & Hypothesis (H)	Ground truth
P: People listening to a choir in a catholic church. **H**: Choir singing in mass.	Entailment
P: A big brown dog swims towards the camera. **H**: A dog is chasing a stick.	Contradiction
P: People waiting at a light on bikes. **H**: There is traffic next to the bikes.	Neutral
P: Three bikers stop in town. **H**: Three bicyclists are riding in a pack.	Neutral

Bi-encoder[3] separately takes two input texts and encodes them into representation vectors, u and v. After obtaining the representation vectors for the two sentences, they are concatenated and fed into a classifier to predict the potential relationship between the two texts (entailment, neutral, or contradiction). Cross-encoder[4] models the NLI task as a pair-wise sentence scoring task by simultaneously taking the premise sentence and the hypothesis sentence as the input to an encoder layer simultaneously, performing full attention over all tokens of the input pair. A Cross-encoder network does not produce sentence embeddings but makes three output values between 0 and 1 for an NLI task and predicts three classes through an argmax operation. Different NLI models usually improve their encoder or embedding layers to achieve higher performance levels of NLI continuously. After the advent of Pre-trained Language Models (PLMs) (e.g., Bidirectional Encoder Representation from Transformers (BERT)[5]), the Bi-encoders and Cross-encoders models integrated with PLMs as their encoder or embedding layers achieve the state-of-the-art performance on most NLI datasets[3, 4].

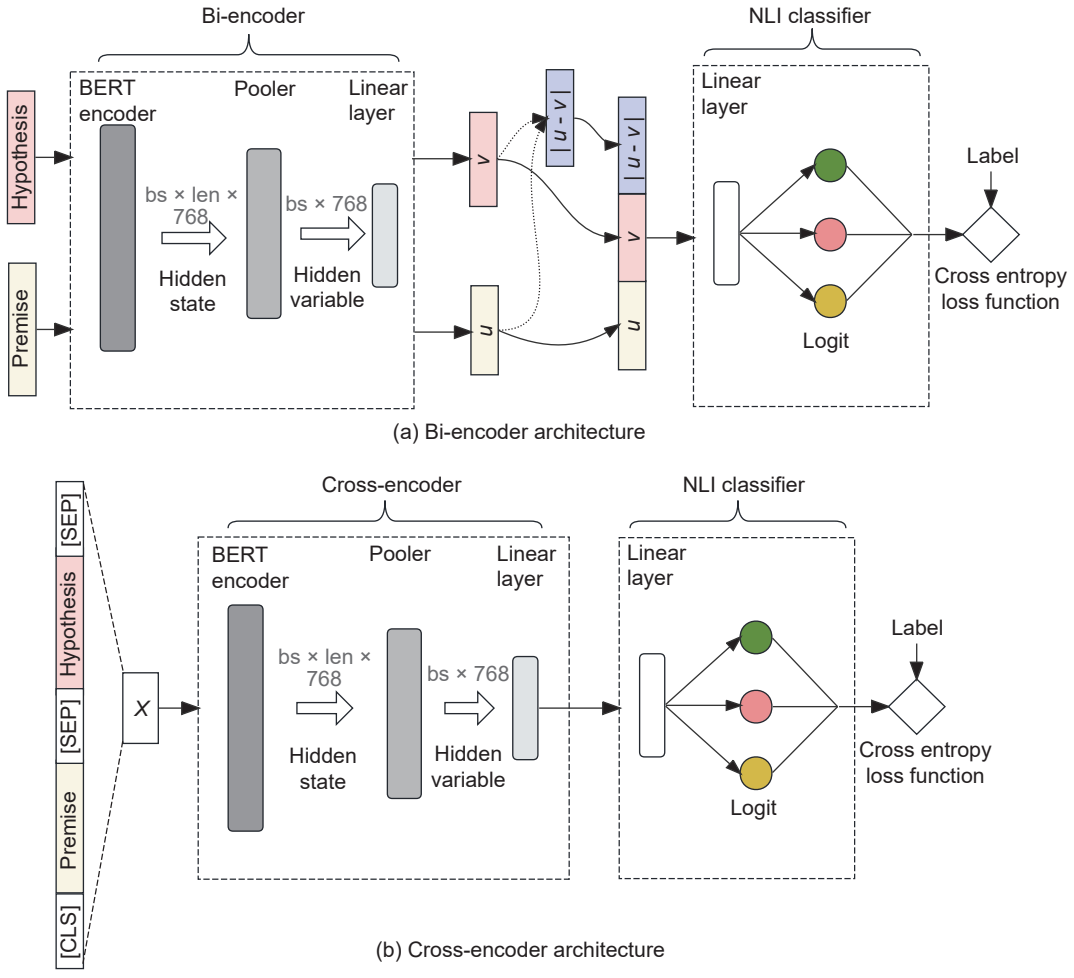


Fig. 1 Bi-encoder and the Cross-encoder architectures, which adopt a variety of pre-trained BERT models as the encoder, set the former state-of-the-art performance on NLI. Specifically, u and v are the vectors obtained by encoding the premise and the hypothesis, respectively, "bs" denotes the batch size, and "len" denotes the length.

Table 2 Six types of bad case statistics from a baseline model in SNLI dataset, predicting the reasoning relationship between the premise and the hypothesis, which shows that most bad cases come from the judgment on the neutral relation between the premise and hypothesis, highlighted in bold.

Ground truth	Wrong prediction	Error proportion (number)
Entailment	**Neutral**	**27.6%** (14 101)
Contradiction	**Neutral**	**20.7%** (10 542)
Neutral	Entailment	**29.7%** (15 129)
Neutral	Contradiction	**15.7%** (8021)
Entailment	Contradiction	4.2% (2134)
Contradiction	Entailment	6.4% (3257)

While existing methods have been successful, we have found that these powerful methods still have a certain proportion of prediction failures when performing NLI tasks. We analyze the cases of prediction failures for different classification results in the NLI task and conduct statistical analysis as shown in Table 2. In Table 2, "Error proportion" gives the proportions of the different types of bad cases in 51 012 total erroneous predictions, which shows that most bad cases come from the judgment on the neutral relation between the premise and hypothesis. This is because the reasoning relationship from the premise to the hypothesis sentences in neutral relationships is very weak and difficult for models to capture, thus leading to errors in predicting NLI relationships.

After studying the counterexamples generated by existing models, we have reconsidered the fundamental features of NLI. Our research reveals two essential characteristics of reasoning when classifying NLI sentences into three categories: (1) When the basic contextual factors of two NLI sentences are unrelated, there is no mutual inference relationship between the two sentences, and this constitutes a Neutral relationship; (2) If the basic contextual factors of two NLI sentences are related, then the two sentences exhibit a non-neutral inference relationship (entailment or contradiction). At this point, the specific classification of the inference relationship (entailment or contradiction) is determined by the reasoning factors within the premises and hypotheses, independent of the basic contextual factors. Meanwhile, the reasoning process is directional, rather than the typical symmetrical text matching process.

We have found that existing methods lack consideration for these two aspects when capturing the reasoning relationship from premises to hypotheses in NLI, as described in the following examples.

Firstly, we noticed that when two NLI sentence pairs have different basic contextual factors, they are essentially semantically unrelated, so their relationship tends to be neutral. However, existing NLI models struggle to disentangle and extract the basic contextual factors from the representation of reasoning factors. For example, in Table 1, the sentences "Three bikers stop in town" and "Three bicyclists are riding in a pack" describe two relatively independent and different basic contextual scenes: the first describes the geographical location of the cyclist, while the second describes their riding style. Therefore, the ground truth label indicating the relationship between them is "Neutral".

However, in actual experiments, it is found that some important baseline models (such as BERT) mistakenly classify the relationship between these two sentences as "contradictory". After analysis, we believe this is because, based on the training methods of these baseline models, they determine the relationship between the two sentences in an undirected matching manner simply. As a result, since the verbs "stop" and "riding" generally do not co-occur or happen simultaneously, these models consider the matching degree between the sentences to be very low, and thus are highly likely to classify their relationship as "contradictory".

In fact, the true reasoning relationship between the two sentences mentioned above is neutral. "Neutral" means that neither sentence can infer whether the other is likely to occur. The reason lies in the fact that the basic contextual factors these two sentences aim to express are unrelated: the first describes the geographical location of the cyclist, while the second describes their riding style. As we can see, the information conveyed by "geographical location" and "riding style" is inferentially unrelated. Therefore, there is no clear reasoning or causal relationship between them—the ground truth label is "neutral".

Therefore, the co-occurrence-based relationship judgment method used in the past is different from the inference-based judgment method required for natural language inference tasks. As a result, the judgments made using these two different approaches may also yield different outcomes. The performance decline in previous models on the same samples is precisely due

to the fact that the co-occurrence-based relationship judgment methods lacked the more refined process of disentangling and extracting reasoning factors that significantly influence classification results from the contextual factors unrelated to classification results.

Secondly, due to the fact that basic contextual factors have little influence on the classification output in the NLI task, when two NLI sentences share similar basic contextual backgrounds, their relationship tends to be non-neutral due to semantic relevance. At this point, further identifying the reasoning factors between them is crucial for correctly determining whether one sentence entail or contradict the other one. We note that the NLI task involves judging a directed reasoning relationship from premise to hypothesis[6]. For example, in Table 1, for the sentences "People listening to a choir in a catholic church" and "Choir singing in mass", some traditional symmetric matching methods can only determine that these two sentences are semantically related. However, because these models often judge that the two core words "listen" and "sing" have a certain degree of mutual exclusivity, they incorrectly judge the relationship between these two sentences as "contradictory". Thus, traditional symmetric matching methods often cannot further determine the specific reasoning relationship between two sentences. In fact, we argue that NLI is not simply an undirectional text matching task but a directional reasoning task. In contrast, humans can correctly infer that the relationship between these two sentences is one of entailment because they can recognize "listen" and "sing" as a directional reasoning process and attribute "sing" and "listen" to a logical sequential process with a before-and-after relationship. Recognizing this point is actually beneficial for discovering and extracting the factors of entailment reasoning. Therefore, if the directional characteristics of the NLI reasoning process are not considered, using traditional symmetric text matching methods may not easily further improve the performance of NLI.

Considering the asymmetry of reasoning in the NLI task and the importance of extracting reasoning factors, we propose a Directed NLI (DNLI) model based on the Cross-encoder and beta-Variational AutoEncoders (VAE) framework for disentangling reasoning factors. As illustrated by the examples in Table 1, reasoning factors, marked with colors, determine the classification results of NLI relationships. In this table, green words denote Entailment reasoning relationships between premises and hypotheses under similar basic contextual conditions. Red words indicate the absence of a reasoning relationship (Contradiction) under similar basic contextual conditions. Yellow words denote that the basic contexts of the preceding and following sentences are unrelated or inconsistent, meaning there is no inference relationship (Neutral). All diagrams in this paper use the same color definitions mentioned above. Firstly, DNLI processes NLI text data in ordered sentence pairs. To achieve this, DNLI employs the Cross-encoder architecture of pre-trained models as the encoder layer, generating encoded representations of text pairs in the latent space. Since the Cross-encoder utilizes special tokens provided by pre-trained models as separators, the order characteristics of the input NLI text pairs are preserved in the generated latent representations. This is advantageous for the model to learn inference relationships, such as "entailment" between preceding and following sentences, instead of erroneously predicting them by treating the semantics of the preceding and following sentences separately. Secondly, DNLI interpretsly utilizes beta-VAE[7, 8] to effectively decouple the directional reasoning trend factors and basic contextual factors in premise-hypothesis text pairs of NLI, by maximizing the mutual information between the representation factors of reasoning trends and NLI classification results. This effectively improves the performance of NLI tasks.

The major contributions of this work are described below:

- We propose an effective NLI training architecture that integrates the disentanglement capability based on beta-VAE and the directed learning capability of the Cross-encoder encoder. It can disentangle the directional reasoning factors and basic contextual factors in premise-hypothesis text pairs, thus providing an interpretable and efficient task performance enhancement solution. It is also applicable to other downstream tasks related to the representation of ordered text pairs.
- We introduce an information-theoretic perspective and mutual information between the reasoning-trend factor representations and the NLI classification results, effectively improving the reasoning performance of NLI tasks.
- The experiments show that our proposed method

exhibits significant improvements over various baseline models and achieves state-of-the-art performance on three widely researched NLI datasets.

2 Related Work

2.1 NLI

NLI is generally viewed as a task of sentence pair modeling, and scoring, and reasoning has been a core topic in artificial intelligence[9]. Pair-wise sentence scoring tasks have wide applications in NLP, including, but not limited to, NLI, information retrieval, and question answering. The most popular architecture to match sentence pairs is the "Siamese" structure. This architecture implements a word/sentence embedding layer by embedding models like Word2Vec[10] and GloVe[11]. On top of the embedding layer is a task-specific encoder based on CNN[12, 13], RNN[14], LSTM[15], and BiLSTM[16] (employs the bidirectional LSTM as a base block). Besides, some effective attention networks with the Siamese structure applying alignment on input pairs are put forward, like ESIM[17], DIIN[18], CAFE[19], DRCN[20], RE2[21], and KNSE[22]. In the final layer of the Siamese architecture, a prediction classifier is utilized to process the final representation vector outputted by the encoder, providing the ultimate prediction result.

Recently, BERT[5] and its variants[‡] have set new state-of-the-art performance in many NLP scenarios. Specifically, they have been applied to two text-matching architectures, the Bi-encoder and Cross-encoder, for pairwise text-scoring tasks. These architectures are extensions and simplifications of the Siamese architecture, as the inclusion of pre-trained models like BERT replaces the originally complex embedding and encoding layers in the Siamese models. A typical implementation of the Bi-encoder architecture based on BERT is Sentence-BERT (SBERT)[3]. SBERT independently maps each input sentence to a dense vector space, generating fixed-size sentence embeddings derived from the output hidden states of BERT. By comparing their sentence embeddings using similarity measures such as cosine similarity or Euclidean distance, the similarity between two sentences is scored. Despite the significant success of the Bi-encoder architecture in NLI tasks, being able to predict whether one text matches another, they can not identify directional reasoning relationships. This is because their design is based on the perspective of symmetric text matching, making them insensitive to the asymmetry of NLI data.

In contrast, a typical model integrating the Cross-encoder architecture and the BERT encoder is CBERT[4]. CBERT models the NLI task as a pairwise sentence scoring task by concatenating the premise and hypothesis beforehand and passing them as input to BERT simultaneously, thus applying global cross-attention over all words in the paired text inputs. The Cross-encoder architecture does not generate sentence embeddings but instead produces three output values between 0 and 1 for the NLI task, predicting three classes through an argmax operation. Most research based on the Cross-encoder framework focuses on improving the encoding performance in their methods with pre-trained models. With the emergence of various pre-trained models based on BERT, CBERT integrated with these models sets the former state-of-the-art performance[3, 4, 23], whose architecture is illustrated in Fig. 1b. Although Cross-encoder models like CBERT simultaneously read both input sentences during training, they do not consider deeper directional reasoning factors in NLI data and can not distinguish confounding factors from the basic context, thereby hindering the performance improvement of classification results.

2.2 Disentangled representation learning for NLP

Disentangled Representation Learning (DRL) is an important approach to achieving better representation learning[24]. DRL aims to acquire a representation where the variation of one dimension is associated with a change in a specific data feature while the variation of other data features remains relatively stable. This characteristic promotes independence among latent factors in the representation and is crucial for achieving disentangled representation learning. Most DRL methods are based on generative models, such as VAE[25] and its variants[7, 26–30]. DRL models have achieved significant success in various domains, including computer vision[31–35], recommendation systems[36–38], graph learning[39], and time-series tasks[40–42].

In the NLP area, DRL has been used more and more. SAVED[43] disentangles the informational content of a tweet from how the information is written and

[‡] https://huggingface.co/models

generates open-domain conversational responses in the desired style. S^2DM[44], a VAE-based disentanglement model, is proposed for multilingual machine reading comprehension by disassociating semantics from syntax. DC-Match[45] disentangles keywords from intents in text semantic matching. CNTM[46] shows that neural topic models have disentanglement ability because they are derived from VAE. Recently, disentanglement representation learning has also been applied to long multi-party dialogue[47] and neural information retrieval[48].

In this paper, we employ disentangled representation learning to decouple latent reasoning factors from the basic context of NLI data. This serves as a prerequisite for optimizing NLI reasoning performance using mutual information maximization.

2.3 Mutual Information (MI) estimation

With latent reasoning factors and basic contextual representations disentangled, we need to maximize the association between reasoning trend factors and NLI classification results. To achieve this, we adopt the MI maximization method.

Recent research enables differentiable and tractable estimation of the MI by deep learning and variational bounds, such as the Donsker-Varadhan[49] divergence, the Jensen-Shannon MI estimator[50], and InfoNCE[51]. In other words, MI can quantify the dependence of two random variables and measure non-linear statistical dependencies between two random variables, generally defined as

$$\mathcal{I}(X;Y) := H(X) - H(X \mid Y) \tag{1}$$

where X and Y denote the input and output of a network. H is the Shannon entropy, and $H(X \mid Y)$ is the conditional entropy of X given Y. Belghazi et al.[52] implemented MI estimation in high-dimensional and continuous scenarios, effectively computing MI between high-dimensional input/output data of deep neural networks. Hjelm et al.[53] proposed Deep InfoMax (DIM), utilizing MI maximization to optimize image data's global and local information representations. Recent research on MI focuses on optimizing a tractable bound. Poole at al.[54] unified methods for estimating upper and lower bounds on MI.

Inspired by the theory of MI, after disentangling the latent reasoning trends within NLI sentence pairs, we employ deep MI optimization to enhance the reasoning effect crucial for improving the classification performance of NLI.

3 Our Approach

This section introduces the proposed model, DNLI. As discussed in Section 1, to achieve interpretable classification of inference results and enhance the performance of NLI tasks, our proposed model disentangles and extracts directional reasoning factors from the latent representations of paired texts, and combines mutual information maximization methods to optimize the representation of reasoning factors that significantly affect classification results. We describe the optimization objectives related to the above two aspects in Section 3.1 and Section 3.2, respectively. In Section 3.3, after incorporating the original classification optimization objectives of the NLI task, the overall optimization objective of DNLI is proposed.

3.1 Disentanglement optimization for asymmetric text pairs

In this section, we present two key aspects as the prerequisites of our interpretable DNLI model: the order-reserved representations of the NLI text pairs and the disentanglement ability. Because of NLI's data asymmetry and considering NLI is essentially a directional relational inference task, disentangling the latent factors of the directional text pair representations is the key to extracting and capturing the directional reasoning factor from the premise to the hypothesis. So, our proposed learning framework for NLI should integrate a directional text pair learning manner and have disentanglement capability.

In our approach, we first leverage BERT's Next-Sentence Prediction (NSP) training manner[5] that reserves the directional relationship features of the text pairs in its encoded text-pair representations. We select the transformer encoder of a pre-trained model, such as BERT[§], to build the NLI data representations because the NSP training method originating from BERT is sensitive to the direction of each input pair, and thus satisfies the first prerequisite that ensures the ordered representations of text pairs for the later extraction of reasoning-trend factors. In training, we concatenate the premise and hypothesis as "[CLS] Premise [SEP] Hypothesis [SEP]" before the formatted pair is handled by the tokenizer of the selected pre-trained model,

[§] Other BERT-derived pre-trained models are available for our framework. We select the base model, BERT (uncased) (https://huggingface.co/google-bert/bert-base-uncased), in our experiments, following the conventions in the literature to ensure fair comparisons.

where "[CLS]" and "[SEP]" are used as the head token and the separator token in a text pair input. We use the last hidden state tokens of the transformer encoder's output as the representations of NLI text pairs. The transformer encoder is used as the encoder layer of our VAE-based disentanglement framework.

Secondly, to satisfy the other prerequisite for the disentanglement ability, we adopt the VAE-based disentanglement architecture and use the pre-trained model's encoder layer as the VAE's encoder network, aiming to learn the independent reasoning-trend factors in NLI data. In detail, we employ the beta-VAE[7, 8] model as our backbone network, which is inherently capable of learning latent disentanglement representations. It allows the disentanglement of the latent reasoning factors from other basic contexts of the premise-hypothesis text pairs (described in Section 3.2 later). Figure 2 illustrates the overall model architecture of our proposed DNLI.

As described in the Part (1) of Fig. 2, VAE is a generative model with inference and generative phases[25]. In its inference phase $q_\phi(z \mid x)$, VAE maps the input X to be latent variables z by an encoder network, which is parameterized with the variational parameters ϕ and usually is modeled as a prior Gaussian distribution $p(z)$. Because the posterior calculation is intractable, the original literature introduces a reparameterization technique to approximate the encoding process with the approximation posterior $q_\phi(z|x)$, which is usually set as a distribution of $N(\mu_\phi(x), \sigma^2_\phi(x))$, where μ represents the mean, σ^2 represents the variance, and ϕ denotes the parameterized encoder network. The reparameterization technique allows sampling z from a normal distribution $N(\mu_\phi(x), \sigma^2_\phi(x))$ by first sampling ϵ from $N(0, I)$ and then computing $z = \mu + \epsilon \times \sigma$.

After that, in the generative phase $p_\theta(x \mid z)$, the reconstruction data X' is decoded from z, where the variational parameter θ denotes the parameterized decoder network. Our model utilizes beta-VAE to alleviate the "Kullback-Leibler (KL) divergence vanishing" (as known as Posterior Collapse) problem existing in the original VAE training[25]. We employ an annealing strategy to dynamically adjust the weight coefficient β of the KL term automatically, ensuring that the VAE network avoids the KL-vanishing during optimization, thereby preventing posterior collapse. The optimization objective of VAE is named Evidence Lower BOund (ELBO) in the VAE's original literature and formulated as below:

$$\mathcal{L}_{\text{ELBO}} = E_{q_\phi}[\log p_\theta(x \mid z)] - \beta D(q_\phi(z \mid x) \| p(z)) \quad (2)$$

which is optimized with the reconstruction/decoder network parameters θ and variational parameters ϕ that parameterize the inference/encoder network.

3.2 MI optimization for NLI classification

Furthermore, to enhance NLI classification effectiveness, DNLI introduces an MI optimization objective aimed at maximizing the MI between NLI classification results and the corresponding latent reasoning factors in text representations, ultimately improving the classification predictive capability of NLI.

As shown in Fig. 2, first, based on the overall latent

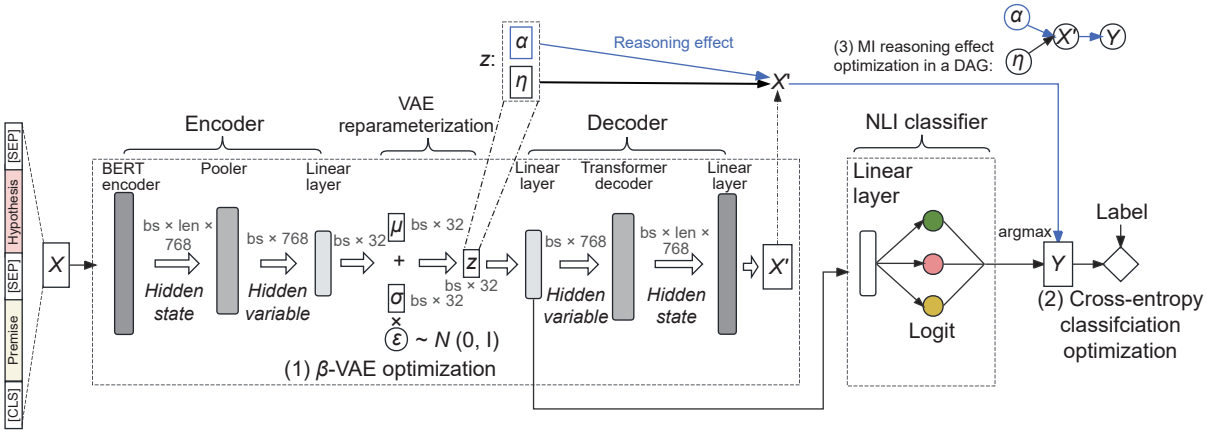


Fig. 2 DNLI architecture. DNLI includes three optimization processes: (1) pair-wise disentanglement representation optimization via beta-VAE, (2) cross-entropy classification optimization for NLI, and (3) MI maximization optimization of reasoning effect in a DAG.

representation z of paired texts, we define the reasoning factor $\alpha \in z$, i.e., those factors in z significantly influence the classifier output Y. These factors can be interpreted as the reasoning-trend factors driving the classification decisions. Simultaneously, we also define the non-reasoning factor $\eta \in z$, representing the shared basic content between the premise and hypothesis. Non-reasoning factors primarily represent content in the text related to the underlying contextual scene, which is only related to the distribution of the reconstructed data X' but does not affect the classification result Y. To describe these two latent factors and their characteristics, we draw on causal effect theory to describe the reasoning effects in the NLI task and introduce a Structural Causal Model (SCM) to explain the interactions between (α, η), X', and Y. This structural causal model is represented as a Directional Acyclic Graph (DAG) in Part (3) of Fig. 2. In the DAG, α and η are mutually independent latent factors in z, learned by the disentanglement module, and interventions on α or η will lead to changes in X', while interventions on α alone will lead to changes in the output classification Y.

Secondly, as previously discussed, we introduce an MI optimization objective, denoted as $\mathcal{L}_{\text{Reasoning}}(\alpha, Y)$, to encourage the influence of reasoning factors in the representation on the classification results. Before delving into this objective, it is crucial to estimate and quantify the reasoning effect from α to Y, represented as $\mathcal{L}(\alpha, Y)$ in the DAG. We adopt causal effect estimation methods to quantify the reasoning effect. Estimating causal effects is an important step in causal inference[55], and various methods for estimating and quantifying causal effects have been proposed in existing literature[56, 57]. However, many of these methods are insufficient for modeling nonlinear neural networks. Fortunately, recent advancements have addressed this limitation by adopting an information-theoretic perspective, employing the MI metric to gauge the causal effect[58]. Inspired by the use of MI estimation to quantify causal effects, our DNLI model also uses MI to compute and quantify the reasoning effects in NLI. MI is a measure traditionally used to quantify the mutual influence between variables. This measure depends on the observed distribution and is the standard definition of conditional mutual information with intervention distributions[59]. Considering two disjoint subsets of internal nodes, denoted as M and N, respectively, the information flow effect from M to N, denoted as $\mathcal{L}(M, N)$, can be computed as follows[58]:

$$\int_M p(m) \int_N p(n \mid \text{do}(m)) \log \frac{p(n \mid \text{do}(m))}{\int_{m'} p(m') p(n \mid \text{do}(m')) \text{d}m'} \text{d}N \text{d}M \tag{3}$$

where $\text{do}(m)$ represents the intervention on the reasoning result under the condition of fixing m[60]. From the above calculation formula, it can be seen that the optimization objective of the reasoning effect from α to Y is consistent with MI optimization. Returning to the natural language inference task of this paper, and considering the mutual independence of (α, η), we similarly utilize the information flow $\mathcal{L}_{\text{Reasoning}}(\alpha, Y)$ to quantify the reasoning effect from α to Y, as shown in the DAG in Fig. 2. The quantification value of the reasoning effect can be obtained by computing the MI defined between α and Y in DNLI using

$$\mathcal{L}_{\text{Reasoning}}(\alpha, Y) = I(\alpha, Y) = E_{\alpha, Y}\left[\log \frac{p(\alpha, Y)}{p(\alpha) p(Y)}\right] \tag{4}$$

which furnishes explanations for our optimization of the NLI classifier from disentanglement and MI perspectives. Particularly considering the disentangled factors $(\alpha, \eta) \in z$, we update the optimization object of the ELBO (see Eq. (2)) as follows:

$$\mathcal{L}_{\text{ELBO}} = E_{q_\phi}[\log p_\theta(x \mid \alpha, \eta)] - \beta D(q_\phi(\alpha, \eta \mid x) \| p(\alpha, \eta)) \tag{5}$$

To elucidate the MI optimization constraint derived from the relationship among latent factors in the presented DAG above, we draw directly upon the findings in Ref. [61] as the foundational theory for our model. Reference [61] provides geometric confirmation of the objective's efficacy in producing explainable factor properties: (1) the reasoning factor accurately captures the direction in which the classifier output changes, and (2) the entire set of latent factors faithfully reflects the data distribution. These pivotal features enhance the interpretability of the original VAE-based disentangled representation learning framework. We designate K and L as hyperparameters representing the number of reasoning and non-reasoning factors, respectively, where $N = K + L$ is the total dimensionality of z.

In summary, our proposed DNLI first disentangles the directional reasoning-trend factors in the representation of text pairs, and then utilizes MI estimation and maximization methods to quantify and optimize the reasoning-trend factor α for their effect on

Y. The overall optimization objective of our DNLI model will be introduced in Section 3.3.

3.3 Optimization objects of DNLI

Our proposed DNLI model employs a fully connected classifier network layer to accomplish NLI classification of paired text representations. As shown in Fig. 2, DNLI takes the hidden variable derived from the VAE decoder network as input to the classifier network layer. It is worth noting that the input to the classifier is based on the reconstructed hidden state of the VAE decoder rather than the output of the encoder. This represents an important architectural difference between DNLI and CBERT. Subsequently, after the latent representation enters the classifier, the model generates prediction probabilities for three classes through a Softmax operation and utilizes standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ to optimize the classification results. The optimization objective of the NLI classifier is expressed as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{i} y_i \log(\hat{y}_i) \tag{6}$$

where y_i is the ground-truth label of sample i and $\hat{y}_i$ is the sample's prediction label obtained through an argmax operation on the predicted probabilities of the three NLI classes ($y_i \in \{0, 1, 2\}$). Without loss of generality, when $y = 0$ and σ signifies a Sigmoid function, the linear separator can be denoted as

$$p(y = 0 \mid x) = \sigma(b^{\text{T}} x) \tag{7}$$

where $b \in \mathbf{R}^N$ signifies the classification decision boundary.

Overall, by integrating the VAE-based disentanglement optimization of latent factors (Eq. (5)), the NLI classifier optimization (Eq. (6)), and the MI optimization on reasoning-trend factors (Eq. (4)), we present the overall optimization objective of DNLI as follows:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\text{ELBO}} + \lambda_2 \cdot \mathcal{L}_{\text{Reasoning}} + \lambda_3 \cdot \mathcal{L}_{\text{CE}} \tag{8}$$

where $\mathcal{L}_{\text{ELBO}}$ generates latent factors for disentanglement, $\mathcal{L}_{\text{Reasoning}}$ further optimizes the classification results using MI maximization of reasoning factors, and $\mathcal{L}_{\text{CE}}$ is the basic NLI classification optimizer. λ_i ($i \in \{1, 2, 3\}$) is the trade-off parameter for the relevant optimization components. In our experiments, their respective values are 0.5, 0.5, and 1. These three optimization functionalities are illustrated in the overall architecture diagram shown in Fig. 2.

4 Experiment

We train DNLI on Nvidia 3090Ti GPUs and implement it by PyTorch. We chose the Adam optimizer, and the learning rate is set to 2×10^{-5}. Sentences in the dataset are all tokenized and converted to lower cases. We also perform a filter on stop words, meaningless symbols and emojis before model training. The maximum sequence length is not limited. Then, we evaluate our proposed DNLI on three well-studied NLI datasets with three class labels. We describe these datasets and the baselines in Sections 4.1 and 4.2. Then, we provide the experimental results and detailed experimental analysis in Section 4.3.

4.1 Datasets

We get the datasets, Stanford Natural Language Inference (SNLI)[62], Multi-genre Natural Language Inference (MNLI), and Adversarial Natural Language Inference (ANLI), through the dataset tool provided by HuggingFace¶ and show their statistics in Table 3. We follow the default splits provided by the datasets and set the batch size as 32/16/32 for SNLI/MNLI/ANLI. We use the original datasets of these three datasets for training and evaluation to ensure a fair comparison with the baseline model.

SNLI☆[62] uses image captions from the Flickr30k dataset as the premise sentences. Its hypothesis

¶ https://huggingface.co/datasets
☆ https://nlp.stanford.edu/projects/snli/

Table 3 Statistics of the datasets. Among the datasets, MNLI features two types of validation sets, while ANLI comprises three versions (r1, r2, and r3).

Dataset	Train split	Validation split	Test split
SNLI	5.50×10^5	1.00×10^4	1.00×10^4
MNLI	4.32×10^5	9.82×10^3 (matched) / 9.83×10^3 (mismatched)	(Without label and not used)
ANLI (r1)	1.69×10^4	1.00×10^3	(Without label and not used)
ANLI (r2)	4.55×10^4	1.00×10^3	(Without label and not used)
ANLI (r3)	1.00×10^5	1.20×10^3	(Without label and not used)

sentences are generated by crowd-sourced annotators who were instructed to judge and manually label the relations of the text pairs as "entailment", "neutral", and "contradiction". SNLI comprises about 5.7×10^5 samples.

MNLI[©][63] is a crowd-sourced dataset offering ten distinct genres of written and spoken English data. Its collection mode is modeled closely like SNLI. The matched test sets originate from the same sources as the MNLI training set, while the mismatched test sets are dissimilar to the MNLI training set. MNLI comprises about 4.33×10^5 sentence pairs.

ANLI[¤][64] is a recently introduced, extensive benchmark dataset for NLI tasks and includes three versions: r1, r2, and r3. It is curated using an iterative and adversarial process involving human annotators and models. The dataset is designed to challenge state-of-the-art models, such as BERT and DeBERTa[65], by selecting particularly difficult examples for these models to classify accurately.

4.2 Baselines

We include seven strong baselines in NLI experiments. We follow the same data pre-processing rules and use the original implementation methods of baselines. We did twenty experiments for each baseline on three datasets to observe the average results of those experiments and listed these results in Table 4.

- **ESIM[66]** is a former state-of-the-art solution for NLI. It operates as a sequential model, designing sequential inference models based on chain LSTMs to deduce contextual relations of two texts.
- **RE2[21]** is an advanced neural interaction network that extracts effective alignment information in text matching. The model is typically suitable for bidirectional matching tasks.
- **BERT[5]** is the bidirectional masked language model. As the most renowned and fundamental pre-trained model at present, to distinguish it from its variants, its configuration is often denoted as "BERT-base-uncased". Its corresponding pre-trained file can be accessed online[◊], with which we vectorize text pairs separately under the Bi-encoder architecture as shown in Fig. 1a and calculate the cosine similarity score between their corresponding vectors for the prediction of NLI.
- **CBERT[4]** includes the process of combining the two input sentences through concatenation and then applying complete attention across the input. CBERT calculates the similarity score with the Cross-encoder architecture, as illustrated in Fig. 1b.
- **InfoBERT[67]** is a learning framework from an information theoretic perspective to perform robust fine-tuning over pre-trained language models.
- **PairSCL[68]** is a supervised contrastive learning approach that operates pair-wise, employing BERT as the encoder. This approach utilizes a cross-attention method to acquire shared representations for sentence pairs.
- **MultiSCL[69]** is an NLI model for low-resource scenarios that employs multi-level supervised contrastive learning. It utilizes a contrastive learning objective at both sentence and pair levels to

© https://cims.nyu.edu/ sbowman/multinli/
¤ https://github.com/facebookresearch/anli
◊ https://huggingface.co/models

Table 4 Accuracy results on all test sets of three datasets. We train our model with the training sets from all three datasets (SNLI + MNLI + ANLI) and report the accuracy scores on the test sets of SNLI, MNLI, and ANLI, respectively. We highlight the best/second-best results in bold/underlined styles.

(%)

Model	SNLI	MNLI			ANLI				Average
		Matched	Mismatched	Average	r1	r2	r3	Average	
ESIM	86.7	72.4	71.9	72.2	48.2	39.4	36.3	41.3	59.2
RE2	88.9	78.8	77.8	78.3	50.2	40.1	37.1	42.4	62.2
BERT	88.1	89.4	89.0	89.2	50.1	38.4	35.0	41.2	65.0
CBERT	92.1	90.5	90.4	<u>90.5</u>	70.5	45.9	42.0	52.8	71.9
InfoBERT	<u>92.2</u>	87.2	87.2	87.2	74.7	50.2	48.5	<u>57.8</u>	<u>73.3</u>
PairSCL	91.9	84.7	77.9	81.3	58.7	44.2	41.0	48.0	66.4
MultiSCL	92.2	87.1	85.4	86.3	66.5	47.7	44.7	53.0	70.6
DNLI (ours)	**93.1**	91.2	90.9	**91.1**	75.3	51.0	49.2	**58.5**	**75.1**

differentiate among various classes of text pairs.

Note that in this paper, we do not consider recent powerful data-augmentation methods (e.g., CounterAL[70]) and label-augmentation methods (e.g., T5-3B[71], based on prompt learning), which effectively improve the evaluation performance of NLI tasks. But to ensure a fair performance comparison with other baseline models under the same conditions on the official datasets, we exclude these methods based on data or label augmentation from our baseline models. We train the selected baselines and our model with the training sets from all three datasets and report the accuracy scores on the test sets, following the training settings of the current state-of-the-art baseline model, InfoBERT.

4.3 Experimental results and analysis

4.3.1 Effectiveness analysis

Table 4 shows the experimental results of our proposed DNLI with other published baseline models on the SNLI/MNLI/ANLI dataset. Especially, BERT, CBERT, InfoBERT, and our model (DNLI) adopt the common pre-trained model, BERT-base[£]. Compared to these former state-of-the-art NLI models, our approach improves the performance by 0.9%, 0.6%, 0.7%, and 1.8% in average accuracy on SNLI, MNLI, ANLI (r1/r2/r3), and the overall average performance across the three datasets, respectively, which indicates that our proposed model is effective in the common NLI tasks.

Our model ensures generalization ability in two ways. (1) According to the reference Eq. (8), the optimization objective of our proposed model consists of three components. From a traditional machine learning perspective, this overall optimization objective is equivalent to adding regularization terms to the single cross-entropy optimization objective of the baseline models. These regularization terms are used to penalize overfitting, thus mitigating potential overfitting issues in baseline models like BERT. Therefore, our model theoretically has better generalization capability compared to the baseline models. (2) Following the standard practice of the baseline model, the training set used for training the model is the union of the training sets from the three datasets as described in Table 4. This approach also avoids the potential reduction in generalization ability that may arise from training the model on the training set of a single dataset.

We report the training speeds of the baseline models (BERT and CBERT) and the proposed DNLI model on an NVIDIA 3090Ti GPU. As shown in Table 5, the BERT model trains more slowly than the Cross-encoder-based CBERT model. This is because BERT generates sentence embeddings for each sentence separately, while CBERT processes two input sentences together without generating individual embeddings, resulting in faster training. Additionally, our proposed DNLI method, which is based on CBERT, does not negatively impact training speed compared to CBERT, although our DNLI model's training time increases due to the VAE-based disentanglement optimization and the MI optimization for reasoning. The additional time mainly comes from the VAE process that fine-tunes the initial embedding layer and its disentanglement optimizations. However, our model converges quickly with the disentanglement optimizations (see Fig. 3).

We note that the maximum batch size setting is limited by the 24 GB memory of the NVIDIA 3090 Ti. For example, the upper limit for the batch size for SNLI is 32 in our hardware environment.

4.3.2 Interpretability analysis

We demonstrate the role of basic contextual factors and reasoning factors in the DNLI model, focusing on the model's interpretability. To this end, we designed a latent variable-based data reconstruction process (the VAE decoder part) in the experiment to reveal the true nature of the two latent factors learned by the DNLI.

As illustrated in Fig. 4, DNLI disentangles the reasoning factors and basic contextual factors in NLI text pairs, enhancing the model interpretability. Only the latent representations relevant to α can affect the classification results: changing the reasoning factor α produces varying reasoning effects in the reconstruction text pair, but changing the non-reasoning factor η affects only the reconstruction basic context that does not relate to the classifier output.

Table 5 Average training speed (iteration number/s) when the learning rate is set to 2×10^{-5}, and the batch size is set to 32/16/32 for SNLI/MNLI/ANLI. It is the optimal combination used in our experiments.

Model	SNLI	MNLI	ANLI
BERT	12.06	16.26	9.22
CBERT	11.88	14.02	8.17
DNLI (ours)	13.85	17.85	10.60

[£] https://huggingface.co/google-bert/bert-base-uncased

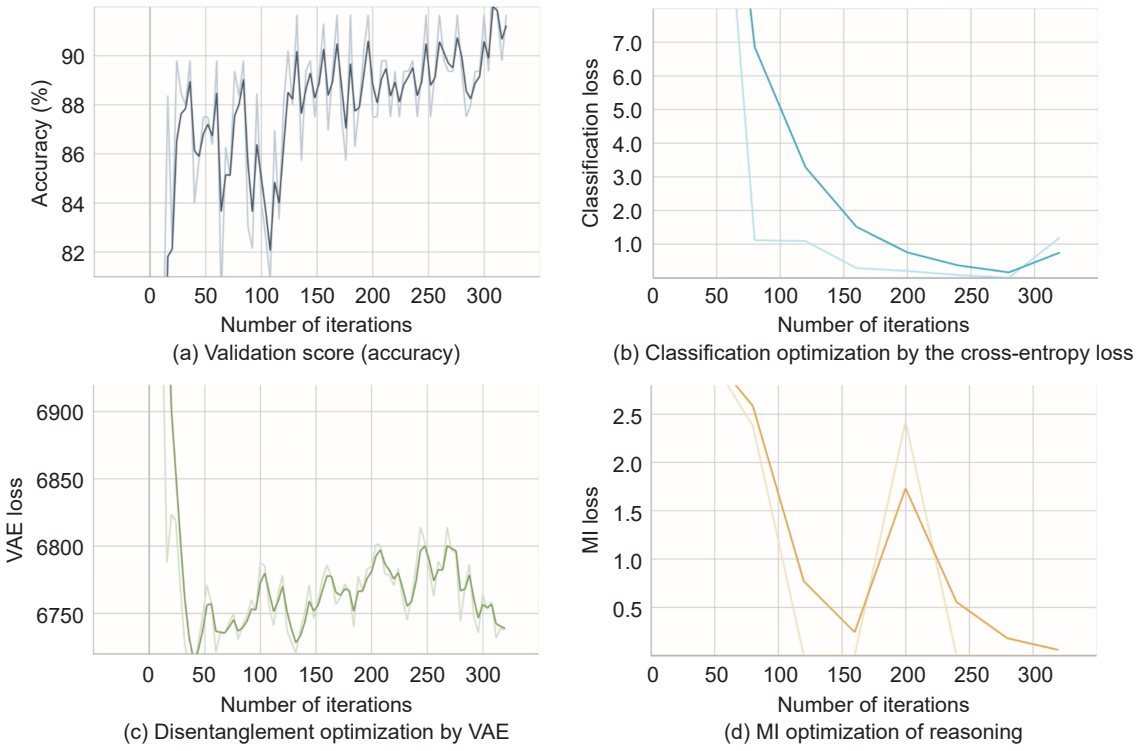


Fig. 3 Accuracy changes during the training of DNLI, along with the convergence of the loss values corresponding to the three optimization processes—classification, disentanglement, and reasoning.

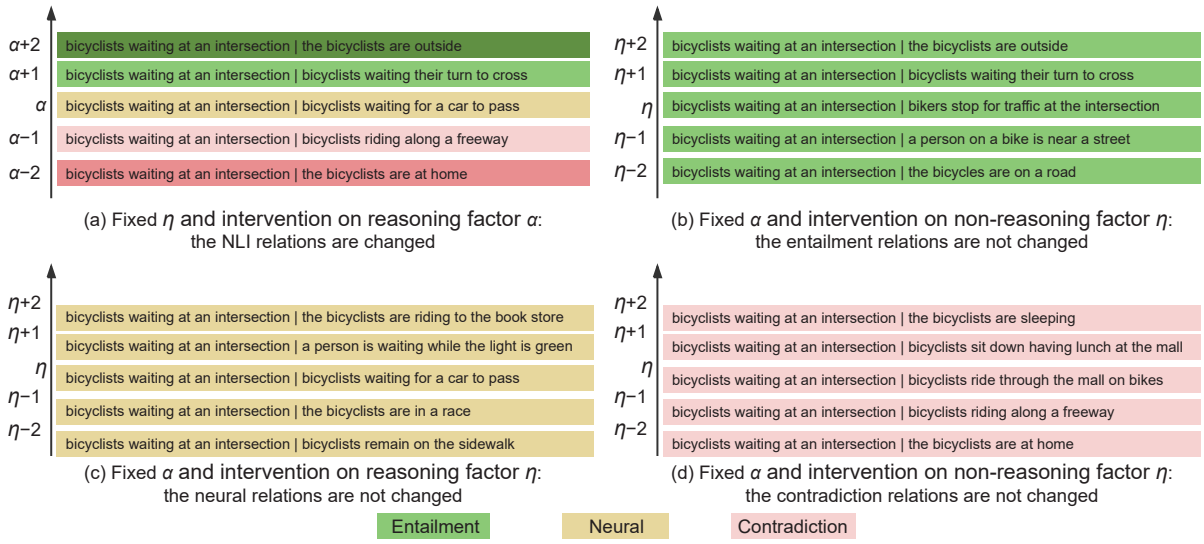


Fig. 4 Illustrations of the effect of altering the learned latent factors. Changing the learned reasoning factor (α) affects the NLI reasoning factors related to the classification results of the text pairs, as shown in (a). Changing the non-reasoning factors (η) affects the basic contexts but does not affect the classification results, as shown in (b), (c), and (d).

With the above experimental results, we can draw the following conclusions related to the interpretability of our model.

Based on the interpretable analysis, we explain the necessity of disentangling the reasoning factors that influence classification results from the contextual semantic factors. By estimating and optimizing reasoning effects when using VAE to optimize the output vectors of the BERT encoder, DNLI distinguishes between the latent factors relevant to the classification result and those irrelevant to it. This is because, without disentanglement optimization of α and η, the VAE reconstruction phase tends to incorporate semantic information from the latent

variables, which may affect classification performance. Therefore, we perform decoupled optimization on α and η, ensuring that more of the semantic contextual information of the original data (η) is preserved in the reconstructed data. At the same time, α retains primarily the information directly related to the classification result. This approach yields better classification performance than directly using BERT encoding outputs. In summary, our model improves the semantic quality of representations through the optimization above while preserving reasoning trend information in representations to achieve better classification results. DNLI simultaneously enhances the classification performance and reconstruction interpretability of representations.

Additionally, we only validate the proposed reasoning disentanglement method on the SNLI, MNLI, and ANLI datasets. This is because, among the current mainstream NLI datasets, only these three datasets are three-class classification datasets. They provide a better foundation for evaluating the effectiveness of our model in reconstructing data across multiple classes (as illustrated in Fig. 4), allowing for a more comprehensive validation of the interpretability of DNLI.

Note that some literature has introduced NLI datasets in low-resource languages, as well as domain-specific NLI datasets recently[72–74]. Since these new datasets are mostly used in relatively specialized environments, we have still used the three most widely used general-purpose datasets in the NLI field in this paper to validate our proposed DNLI model and other baseline models.

Overall, the practical significance of our proposed DNLI is as follows. First, DNLI combines disentangled representation learning with pre-trained models, not only improving the performance of current NLI tasks, but also enhancing the interpretability of the representation vectors generated by pre-trained models. As shown in Fig. 4, the disentangled features, or representation factors, obtained through optimization, allow for control over specific attributes in text generation—such as inference relations, topics, and sentiments—during text reconstruction. This makes it possible to intuitively understand and control the process for specific tasks, which will benefit the interpretability required for the future generative tasks of large language models. Secondly, the independent semantic feature representation factors obtained through DNLI's disentanglement process help reduce reliance on labeled data in unsupervised or weakly supervised conditions. While improving the accuracy of relation classification in the current task, our method can also effectively generalize to other classification tasks.

4.3.3 Ablation study

Our model exhibits two primary enhancements when contrasted with another former state-of-the-art model, CBERT, as shown in Fig. 1b. The first is incorporating VAE framework optimization, as illustrated in Part (1) in Fig. 2, while the second involves the integration of reasoning effect optimization, as shown in Part (3) in Fig. 2. We run ablation tests to understand better the contribution of two primary enhancements from our proposed DNLI model. The results are shown in Table 6.

Based on the results of ablation studies, we can further explain the performance differences between the sentence representations based on VAE used in DNLI and the sentence representations directly obtained from BERT (or CBERT) outputs during classification training. When only using the output vectors of the BERT encoder for classification training, the resulting sentence representations tend to gradually lose the intrinsic semantics of the sentences, as they primarily retain information relevant to the classification result. Conversely, by applying the VAE architecture to optimize the output vectors of the BERT encoder for reconstruction, it encourages the learned representations to preserve the semantic information of

Table 6 Ablation study of our proposed DNLI. The first setting of "DNLI without VAE & MI" is equivalent to the baseline CBERT of the Cross-encoder architecture, as shown in Fig. 1b. We report the accuracy results.

(%)

Ablation setting	SNLI	MNLI		ANLI		
		Matched	Mismatched	r1	r2	r3
DNLI without VAE & MI	92.2	90.5	90.4	70.5	45.9	42.0
DNLI without MI	92.5	91.0	90.6	71.9	47.4	44.8
DNLI	93.1	91.2	90.9	75.3	51.0	49.2

the original data to the fullest extent possible.

4.3.4 Parameter analysis

We search for different hyper-parameter options and choose the optimal hyper-parameter value for our model. We give the parameter study for the impact of the number of reasoning-trend factors (K). As shown in Table 7, we choose the best results by tuning K on different datasets. From the experiment results on the number of reasoning-trend factors, we can infer that due to the different distribution characteristics of different datasets, we can select appropriate K values for different datasets to accurately separate and store latent factors related to the NLI classification results, thereby achieving the performance improvement for NLI tasks by tuning the K value.

4.3.5 Limitations and future work

Some recent research[71, 75] has found the problem of overestimating NLI model performance due to superficial correlation biases from the NLI datasets. For example, a crowd worker who makes NLI datasets (e.g., SNLI) usually generates a contradiction hypothesis to contradict the information in the premise by simply introducing negation words (e.g., not, nothing, or nobody), which enables some existing NLI models to directly predict the relationship labels of the hypotheses based on the negation words, even without considering the input premise. To verify the overestimated evaluation of the performance of the existing NLI models because of the dataset bias, authors of some important NLI datasets (SNLI and MNLI) partition each NLI test set into two subsets: examples that can be accurately predicted without considering the premise are labeled "Easy", and those it could not are labeled "Hard". When testing the Hard versions of the NLI datasets, many existing NLI models have decreased performance.

We believe our proposed DNLI can alleviate the superficial correlation biases in its training process because our model introduces VAE and its reconstruction optimization, which ensures that sufficient semantic information is retained in the latent variables for better reconstructing the data, thereby alleviating the problem of only superficial correlation bias information in the learned latent representations. However, as unbiased optimization is not the focus of this paper and can be looked at as an extension advantage of our model, we validate our model on the full versions of the NLI datasets without adding and validating their "Hard" versions, which can be in future work.

5 Conclusion

Leveraging the data asymmetry inherent in NLI data, we present a simple and effective natural language inference model, DNLI, to capture latent directional reasoning factors in paired texts, from premises to hypotheses. DNLI employs BERT (uncased) as the encoder layer to preserve the sequential information of input text pairs in text representations. Subsequently, it adopts the VAE framework into the pair-wise text-matching architecture to disentangle the reasoning factors in paired text and finally enhances NLI classification performance through MI maximization. Consequently, we obtain interpretable text representations and state-of-the-art classification performance on three widely studied NLI datasets, SNLI, MNLI, and ANLI.

Acknowledgment

This work was supported by the National Key R&D Program of China (No. 2022ZD0160703), the National Natural Science Foundation of China (Nos. 62202422 and 62071330), the Natural Science Foundation of Shandong Province (No. ZR2021MH227), and the Shanghai Artificial Intelligence Laboratory.

Table 7 Experimental setup for the hyper-parameter (number of reasoning-trend factors K) tuning for our proposed DNLI when we set the dimension number of the latent variable z as 32. We report the accuracy results (the best are in bold).

(%)

K	SNLI	MNLI (mached)	MNLI (mismatched)	ANLI (r1)	ANLI (r2)	ANLI (r3)
1	92.3	90.3	90.2	75.0	51.0	45.9
2	92.8	90.7	90.4	75.2	**51.0**	48.5
3	**93.1**	91.0	90.7	75.1	50.3	**49.2**
4	93.0	**91.2**	90.6	**75.3**	49.6	48.8
5	92.4	90.9	**90.9**	74.5	49.3	48.2
6	91.8	90.3	90.8	74.4	49.6	46.5

References

[1] B. MacCartney and C. D. Manning, Modeling semantic containment and exclusion in natural language inference, in *Proc. 22nd Int. Conf. Computational Linguistics* (*Coling 2008*), Manchester, UK, 2008, pp. 521–528.

[2] J. Mueller and A. Thyagarajan, Siamese recurrent architectures for learning sentence similarity, in *Proc. 30th AAAI Conf. Artificial Intelligence*, Phoenix, AZ, USA, 2016, pp. 2786–2792.

[3] N. Reimers and I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, China, 2019, pp. 3982–3992.

[4] S. Humeau, K. Shuster, M. A. Lachaux, and J. Weston, Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring, in *Proc. 8th Int. Conf. Learning Representations*, Addis Ababa, Ethiopia, https://openreview.net/forum?id=SkxgnnNFvH, 2020.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies, Volume 1* (*Long and Short Papers*), Minneapolis, MN, USA, 2019, pp. 4171–4186.

[6] I. Dagan, O. Glickman, and B. Magnini, The PASCAL recognising textual entailment challenge, in *Proc. 1st Pascal Machine Learning Challenges Workshop*, Southampton, UK, 2005, pp. 177–190.

[7] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, *β*-VAE: Learning basic visual concepts with a constrained variational framework, in *Proc. 5th Int. Conf. Learning Representations*, Toulon, France, https://openreview.net/forum?id=Sy2fzU9gl, 2017.

[8] R. T. Q. Chen, X. Li, R. Grosse, and D. K. Duvenaud, Isolating sources of disentanglement in VAEs, in *Proc. 32nd Int. Conf. Neural Information Processing Systems*, Montréal, Canada, 2018, pp. 2615–2625.

[9] Z. Zheng and X. Zhu, NatLogAttack: A framework for attacking natural language inference models with natural logic, in *Proc. 61st Annu. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Toronto, Canada, 2023, pp. 9960–9976.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, in *Proc. 27th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2013, pp. 3111–3119.

[11] J. Pennington, R. Socher, and C. Manning, GloVe: Global vectors for word representation, in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, Qatar, 2014, pp. 1532–1543.

[12] B. Hu, Z. Lu, H. Li, and Q. Chen, Convolutional neural network architectures for matching natural language sentences, in *Proc. 28th Int. Conf. Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 2042–2050.

[13] W. Yin and H. Schütze, Convolutional neural network for paraphrase identification, in *Proc. 2015 Conf. North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Denver, CO, USA, 2015, pp. 901–911.

[14] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng, A deep architecture for semantic matching with multiple positional sentence representations, in *Proc. 30th AAAI Conf. Artificial Intelligence*, Phoenix, AZ, USA, 2016, pp. 2835–2841.

[15] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiský, and P. Blunsom, Reasoning about entailment with neural attention, arXiv preprint arXiv: 1509.06664, 2015.

[16] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, Supervised learning of universal sentence representations from natural language inference data, in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 670–680.

[17] L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, and Z. Jin, Natural language inference by tree-based convolution and heuristic matching, in *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016, pp. 130–136.

[18] Y. Gong, H. Luo, and J. Zhang, Natural language inference over interaction space, in *Proc. 6th Int. Conf. Learning Representations*, Vancouver, Canada, https://openreview.net/forum?id=r1dHXnH6-, 2017.

[19] Y. Tay, L. A. Tuan, and S. C. Hui, A compare-propagate architecture with alignment factorization for natural language inference. arXiv preprint arXiv: 1801.00102, 2017.

[20] S. Kim, I. Kang, and N. Kwak, Semantic sentence matching with densely-connected recurrent and co-attentive information, in *Proc. 33rd AAAI Conf. Artificial Intelligence*, Honolulu, HI, USA, 2019, pp. 6586–6593.

[21] R. Yang, J. Zhang, X. Gao, F. Ji, and H. Chen, Simple and effective text matching with richer alignment features, in *Proc. 57th Annu. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 4699–4709.

[22] W. Chen, S. Wei, Z. Wei, and X. Huang, KNSE: A knowledge-aware natural language inference framework for dialogue symptom status recognition, in *Proc. Findings of the Association for Computational Linguistics*: *ACL 2023*, Toronto, Canada, 2023, pp. 10278–10286.

[23] X. Zhou, X. Jie, S. Zhou, K. Shi, Z. Yu, J. Bu, and H. Wang, Inversive-reasoning augmentation for natural language inference, in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*), Seoul, Republic of Korea, 2024, pp. 10351–10355.

[24] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[25] D. P. Kingma and M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv: 1312.6114, 2014.
[26] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, Disentangling disentanglement in variational autoencoders, in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, CA, USA, 2019, pp. 4402–4412.
[27] S. Burkhardt and S. Kramer, Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model, *J. Mach. Learn. Res.*, vol. 20, no. 131, pp. 1–27, 2019.
[28] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, CA, USA, 2019, pp. 4114–4124.
[29] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, CausalVAE: Disentangled representation learning via neural structural causal models, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 9588–9597.
[30] J. Bai, W. Wang, and C. Gomes, Contrastively disentangled sequential variational autoencoder, in *Proc. 35th Int. Conf. Neural Information Processing Systems*, Virtual Event, 2021, p. 773.
[31] F. Trauble, E. Creager, N. Kilbertus, F. Locatello, A. Dittadi, A. Goyal, B. Schölkopf, and S. Bauer, On disentangled representations learned from correlated data, in *Proc. 38th Int. Conf. Machine Learning*, Virtual Event, 2021, pp. 10401–10412.
[32] S. Lee, S. Cho, and S. Im, DRANet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 15247–15256.
[33] Y. Liu, E. Sangineto, Y. Chen, L. Bao, H. Zhang, N. Sebe, B. Lepri, W. Wang, and M. De Nadai, Smoothing the disentangled latent style space for unsupervised image-to-image translation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 10780–10789.
[34] Y. Zhang, Y. Zhang, W. Guo, X. Cai, and X. Yuan, Learning disentangled representation for multimodal cross-domain sentiment analysis, *IEEE Trans. Neural Network. Learn. Syst.*, vol. 34, no. 10, pp. 7956–7966, 2023.
[35] S. Alaniz, M. Federici, and Z. Akata, Compositional mixture representations for vision and text, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, New Orleans, LA, USA, 2022, pp. 4201–4210.
[36] X. Wang, H. Chen, and W. Zhu, Multimodal disentangled representation for recommendation, in *Proc. 2021 IEEE Int. Conf. Multimedia and Expo* (*ICME*), Shenzhen, China, 2021, pp. 1–6.
[37] H. Chen, Y. Chen, X. Wang, R. Xie, R. Wang, F. Xia, and W. Zhu, Curriculum disentangled recommendation with noisy multi-feedback, in *Proc. 35th Int. Conf. Neural Information Processing Systems*, Virtual Event, 2021, p. 2062.
[38] X. Wang, H. Chen, Y. Zhou, J. Ma, and W. Zhu, Disentangled representation learning for recommendation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 408–424, 2023.
[39] H. Li, Z. Zhang, X. Wang, and W. Zhu, Disentangled graph contrastive learning with independence promotion, *IEEE Trans. Knowledge Data Eng.*, vol. 35, no. 8, pp. 7856–7869, 2023.
[40] Y. Li, Z. Chen, D. Zha, M. Du, J. Ni, D. Zhang, H. Chen, and X. Hu, Towards learning disentangled representations for time series, in *Proc. 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, Washington, DC, USA, 2022, pp. 3270–3278.
[41] Z. Wang, X. Xu, W. Zhang, G. Trajcevski, T. Zhong, and F. Zhou, Learning latent seasonal-trend representations for time series forecasting, in *Proc. 36th Int. Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2022, p. 2810.
[42] X. Jie, X. Zhou, C. Su, Z. Zhou, Y. Yuan, J. Bu, and H. Wang, Disentangled anomaly detection for multivariate time series, in *Proc. Companion Proc. ACM Web Conf. 2024*, Singapore, 2024, pp. 931–934.
[43] J. Dougrez-Lewis, M. Liakata, E. Kochkina, and Y. He, Learning disentangled latent topics for Twitter rumour veracity classification, in *Proc. Findings of the Association for Computational Linguistics*: *ACL-IJCNLP 2021*, Virtual Event, 2021, pp. 3902–3908.
[44] L. Wu, S. Wu, X. Zhang, D. Xiong, S. Chen, Z. Zhuang, and Z. Feng, Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension, in *Proc. 60th Annu. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Dublin, Ireland, 2022, pp. 991–1000.
[45] Y. Zou, H. Liu, T. Gui, J. Wang, Q. Zhang, M. Tang, H. Li, and D. Wang, Divide and conquer: Text semantic matching with disentangled keywords and intents, in *Proc. Findings of the Association for Computational Linguistics*: *ACL 2022*, Dublin, Ireland, 2022, pp. 3622–3632.
[46] X. Zhou, J. Bu, S. Zhou, Z. Yu, J. Zhao, and X. Yan, Improving topic disentanglement via contrastive learning, *Inf. Process. Manage.*, vol. 60, no. 2, p. 103164, 2023.
[47] X. Ma, Z. Zhang, and H. Zhao, Structural characterization for dialogue disentanglement, in *Proc. 60th Annu. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Dublin, Ireland, 2022, pp. 285–297.
[48] X. Zhou, Y. Gao, X. Jie, X. Cai, J. Bu, and H. Wang, EASE-DR: Enhanced sentence embeddings for dense retrieval, in *Proc. 47th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Washington, DC, USA, 2024, pp. 2374–2378.
[49] M. D. Donsker and S. R. S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. IV, *Commun. Pure Appl. Math.*, vol. 36, no. 2, pp. 183–212, 1983.
[50] S. Nowozin, B. Cseke, and R. Tomioka, *f*-GAN: Training generative neural samplers using variational divergence minimization, in *Proc. 30th Int. Conf. Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 271–279.

[51] A. van den Oord, Y. Li, and O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv: 1807.03748, 2018.

[52] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, Mutual information neural estimation, in *Proc. 35th Int. Conf. Machine Learning*, Stockholm, Sweden, 2018, pp. 531–540.

[53] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, Learning deep representations by mutual information estimation and maximization, in *Proc. 7th Int. Conf. Learning Representations*, New Orleans, LA, USA, https://openreview.net/forum?id=Bklr3j0cKX, 2019.

[54] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, On variational bounds of mutual information, in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, CA, USA, 2019, pp. 5171–5180.

[55] X. Dong, J. Xu, Y. Bu, C. Zhang, Y. Ding, B. Hu, and Y. Ding, Beyond correlation: Towards matching strategy for causal inference in information science, *J. Inf. Sci.*, vol. 48, no. 6, pp. 735–748, 2022.

[56] R. C. Lewontin, The analysis of variance and the analysis of causes, *Am. J. Hum. Genet.*, vol. 26, no. 3, pp. 400–411, 1974.

[57] P. W. Holland, Causal inference, path analysis and recursive structural equations models, *ETS Res. Rep. Ser.*, vol. 1988, no. 1, pp. 1–50, 1988.

[58] N. Ay and D. Polani, Information flows in causal networks, *Adv. Complex Syst.*, vol. 11, pp. 17–41, 2008.

[59] T. Q. Tran, K. Fukuchi, Y. Akimoto, and J. Sakuma, Unsupervised causal binary concepts discovery with VAE for black-box model explanation, in *Proc. 36th AAAI Conf. Artificial Intelligence*, Virtual Event, 2022, pp. 9614–9622.

[60] J. Pearl, *Causality*: *Models, Reasoning, and Inference*, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

[61] M. O'Shaughnessy, G. Canal, M. Connor, M. Davenport, and C. Rozell, Generative causal explanations of black-box classifiers, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, p. 458.

[62] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, A large annotated corpus for learning natural language inference, in *Proc. 2015 Conf. Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 632–642.

[63] A. Williams, N. Nangia, and S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies, Volume 1* (*Long Papers*), New Orleans, LA, USA, 2018, pp. 1112–1122.

[64] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, Adversarial NLI: A new benchmark for natural language understanding, in *Proc. 58th Annu. Meeting of the Association for Computational Linguistics*, Virtual Event, 2020, pp. 4885–4901.

[65] P. He, X. Liu, J. Gao, and W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, in *Proc. 9th Int. Conf. Learning Representations*, Virtual Event, https://openreview.net/forum?id=XPZIaotutsD, 2021.

[66] Q. Chen, X. Zhu, Z. H. Ling, S. Wei, H. Jiang, and D. Inkpen, Enhanced LSTM for natural language inference, in *Proc. 55th Annu. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Vancouver, Canada, 2017, pp. 1657–1668.

[67] B. Wang, S. Wang, Y. Cheng, Z. Gan, R. Jia, B. Li, and J. Liu, InfoBERT: Improving robustness of language models from an information theoretic perspective, in *Proc. 9th Int. Conf. Learning Representations*, Virtual Event, https://openreview.net/forum?id=hpH98mK5Puk, 2021.

[68] S. Li, X. Hu, L. Lin, and L. Wen, Pair-level supervised contrastive learning for natural language inference, in *Proc. ICASSP 2022 - 2022 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*), Singapore, 2022, pp. 8237–8241.

[69] S. Li, X. Hu, L. Lin, A. Liu, L. Wen, and P. S. Yu, A multi-level supervised contrastive learning framework for low-resource natural language inference, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1771–1783, 2023.

[70] X. Deng, W. Wang, F. Feng, H. Zhang, X. He, and Y. Liao, Counterfactual active learning for out-of-distribution generalization, in *Proc. 61st Annu. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Toronto, Canada, 2023, pp. 11362–11377.

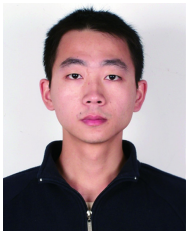
[71] P. Kavumba, A. Brassard, B. Heinzerling, and K. Inui, Prompting for explanations improves adversarial NLI. Is this true? {Yes} it is {true} because {it weakens superficial cues}, in *Proc. Findings of the Association for Computational Linguistics*: *EACL 2023*, Dubrovnik, Croatia, 2023, pp. 2165–2180.

[72] M. Sadat and C. Caragea, MSciNLI: A diverse benchmark for scientific natural language inference, in *Proc. 2024 Conf. North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (*Volume 1*: *Long Papers*), Mexico City, Mexico, 2024, pp. 1610–1629.

[73] S. Mandal and A. Modi, IITK at SemEval-2024 task 2: Exploring the capabilities of LLMs for safe biomedical natural language inference for clinical trials, in *Proc. 18th Int. Workshop on Semantic Evaluation* (*SemEval-2024*), Mexico City, Mexico, 2024, pp. 1397–1404.

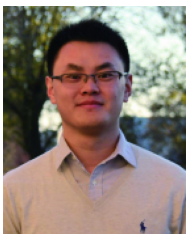
[74] E. Poesina, C. Caragea, and R. Ionescu, A novel cartography-based curriculum learning method applied on RoNLI: The first Romanian natural language inference corpus, in *Proc. 62nd Annu. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Bangkok, Thailand, 2024, pp. 236–253.

[75] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, Annotation artifacts in natural language inference data, in *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies, Volume 2* (*Short Papers*), New Orleans, LA, USA, 2018, pp. 107–112.

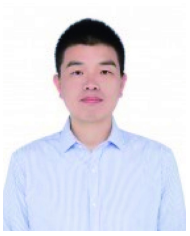


Xixi Zhou received the PhD degree in computer science from Zhejiang University, China in 2024. He is currently collaborating at The First Affiliated Hospital, Zhejiang University School of Medicine, China. His research interests include machine learning, natural language processing, and information systems. He previously worked at Alibaba Group.

Ziping Zhao received the PhD degree from Nankai University, China in 2008. He is currently a professor and the associate dean at College of Computer, Tianjin Normal University, China. His research has primarily focused on speech emotion recognition and speech-based automatic analysis of depression diagnosis.

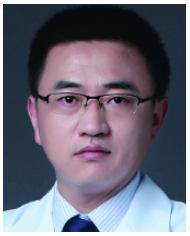


Haishuai Wang received the PhD degree from University of Technology Sydney, Australia in 2017. He is currently a professor at College of Computer Science, Zhejiang University, China. He was selected for the Hundred-Talent Program of Zhejiang University (ZJU100), and was a faculty member at Harvard Medical School and Fairfield University, USA. His research interests include data mining and machine learning.



Limin Zeng received the PhD degree from Technische University Dresden (TU Dresden), Germany in 2013. He is currently an associate professor at Zhejiang University, China. His research focuses primarily on innovative intelligent human-computer interaction technologies, robotics, and vision-based applications using deep learning.

Jiajun Bu received the BEng and PhD degrees in computer science from Zhejiang University, China in 1995 and 2000, respectively. He is currently a professor at College of Computer Science, Zhejiang University, China. His research interests include data mining, information retrieval, medical data analysis, and information accessibility.



Wenjie Liang received the PhD degree from Zhejiang University, China in 2017. He is currently a PhD supervisor at Department of Radiology, The First Affiliated Hospital, Zhejiang University School of Medicine, China. His research interests include medical image analysis and diagnostic methods based on artificial intelligence algorithms.