




Fine-grained cybersecurity entity typing based on multimodal representation learning

BaoLei Wang^{1,2} · Xuan Zhang^{1,3,4}  · JiShu Wang⁵ · Chen Gao⁵ · Qing Duan^{1,3,4} · LinYu Li¹

Received: 7 August 2022 / Revised: 9 June 2023 / Accepted: 31 August 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Fine-grained entity typing is crucial to improving the efficiency of research in the field of cybersecurity. However, modality limitations and type-labeling hierarchy complexity limit the construction of fine-grained entity typing datasets and the performance of related models. Therefore, in this paper, we constructed a fine-grained entity typing dataset based on multimodal information from the cybersecurity literatures and design a multimodal representation learning model based on it. Specifically, we design and introduce a new benchmark dataset called CySets to facilitate the study of new tasks and train a novel multimodal representation learning model called Cyst-MMET with multitask objectives. The model utilizes multimodal knowledge from literature and external to unify visual and textual representations by eliminating visual noise through a multi-level fusion encoder, thereby alleviating data bottlenecks and long-tail problems in the fine-grained entity typing task. Experimental results show that CySets have sharper hierarchies and more diverse labels than the existing datasets. Across all datasets, our model achieves state-of-the-art or dominant performance (3%), demonstrating that the model is effective in predicting entity types at different granularities.

Keywords Cybersecurity · Information extraction · Fine-grained · Multimodal

✉ Xuan Zhang
zhxuan@ynu.edu.cn

¹ School of Software, Yunnan University, Kunming 650091, Yunnan, China

² The Yi-Shu-Si River Basin Administration Hydrological Bureau, HRC, Xuzhou, China

³ Key Laboratory of Software Engineering of Yunnan Province, Kunming 650091, Yunnan, China

⁴ Engineering Research Center of Cyberspace, Kunming 650091, Yunnan, China

⁵ School of Information Science & Engineering, Yunnan University, Kunming 650091, Yunnan, China

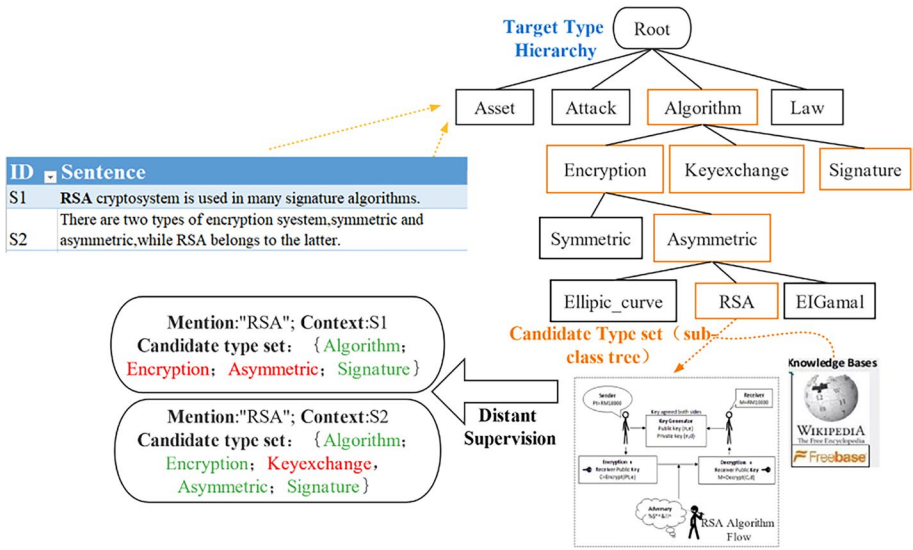


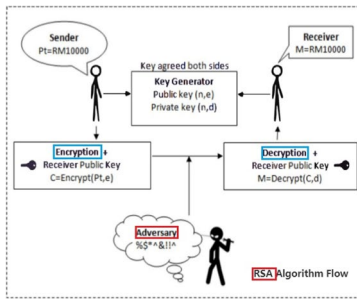
Fig. 1 Entity hierarchy and FET task. Given entity mentions and their contexts, the set of possible type labels is predicted, where the yellow boxes indicate for the mentions and context, the set of candidate entities generated by distant supervision, and green indicates the relatively correct entity labels, while red indicates the wrong labels

1 Introduction

Entity typing, which learns text features and extracts meaningful information for classifying entity classes, is a burgeoning research area in natural language processing (NLP). Entity labeling techniques were utilized in early work on entity typing [1, 2]. To label entities, early work concentrated on coarse-grained entity typing tasks, such as named entity recognition (NER) [3]. However, in recent years, the development of deep learning and neural networks has prompted researchers to shift their attention from coarse-grained to fine-grained entity typing (FET) tasks. According to the given entity mentions and the matching contexts, FET seeks to categorize entity mentions into fine-grained semantic label sets. Considering the example in Fig. 1, for the context “[RSA] cryptosystem is used in many signature algorithms”, the mention of “RSA” in the traditional NER task would type it as *Asymmetric*. However, under a FET scheme, it may be classified as a type label set {/Encryption, Security Algorithm, /Asymmetric cryptosystem}. Many NLP tasks can benefit from providing fine-grained semantic labels, such as entity relationship extraction [4, 5], knowledge base construction and extension [6], entity linking [7], and question and answer [8].

The fundamental challenge of FET comes from data bottlenecks and complex sets of fine-grained entity hierarchy labels, which further limits the performance of the FET model.

First, entity representation is affected by modal limitations. The FET models proposed in recent years [9–11] focus on textual information to accomplish the entity typing task. However, the information provided by a single modality is limited and possibly misleading, which seriously affects the performance of FET models. In particular, the



Text: As shown in the figure, we use traditional cryptosystem as our [signature] algorithm.

(a)



Text: Using Fake DeFi Apps, [he] targeted investors' Bitcoin Assets.

(b)

Fig. 2 Example of multimodal enhanced entity representation

use of distant supervised methods introduces noise labels in the training data, i.e., all possible types of entities are assigned to its mention. However, mentions are free to assume different types depending on their contexts. As shown in Fig. 2a, considering the type hierarchy sampled from the CySets, for the mention of “cryptosystem” in the sentence the traditional FET model can only recognize {*Security Algorithm*, /*Signature*, /*Asymmetric*}, but the latent type /*RSA* cannot be inferred from the context of the mention. It is not inferred from the mentioned context and is therefore considered a “noisy” label. However, by matching the image information with the text, we can find the image region “*RSA Algorithm Flow*” (the red area in Fig. 2a) associated with the text information in the image and thus infer the type label /*RSA*.

It is worth mentioning that FET has faced significant difficulties in modeling the type hierarchy of fine-grained entities in the presence of noisy labels. Previous research has mostly relied on heuristics, for example, Xu [12] employed reinforcement learning models to learn how to describe type hierarchies, while Gillick [13] used a set of heuristics to eliminate noisy training data. However, due to the reduced amount of training data, such methods lead to performance degradation and are difficult to fundamentally enhance the breadth and accuracy of entity recognition. For example, in the text description in Fig. 2b, it is difficult to determine the fine-grained type of the entity [he] because it may describe any character role such as an athlete, an adult, or a teacher. In addition, in the above example, if auxiliary information such as images were missing, the existing model like Ling [16] would assign pronouns such as [he] to the category of “*PER*”. We call this the “long tail problem,” where too many type labels are assigned to one category. This problem can be solved by using multimodal information such as images as additional inputs to enhance the text representation and improve the accuracy and breadth of entity recognition.

Second, entities are represented in a fine-grained manner [14]. In multimodal representation learning models, entities are usually represented by specific words or image regions in the text. However, if recognition methods treat each word and image region equally, simply connecting different modal representations may ignore the interactions between modalities. Lin [10] used a hybrid classification approach beyond binary correlation to exploit the type interdependence of potential type representations. Based on this, Sun [15] used Bi-LSTM to connect chemical structure graph and textual description representations to accomplish the FET task. Unlike its shallow modal interactions, we construct a fusion encoder to align feature

representations of different modalities, going beyond the traditional concept of named entities, with some recognition of pronouns, etc., to solve the problem of modal contradictions due to irrelevant visual noise. As in Fig. 2b, the area of the image where the “person” is located should match the contextual mention [*he*]. The mention [*he*] should, in theory, have a low similarity to other parts of the image and a high resemblance to the area of the image where the “person” is located. However, due to modal heterogeneity, the result may be just the opposite. This results in a semantic gap between different modalities, affecting entity recognition performance.

Finally, building domain FET datasets is challenging. Domain datasets are complex and require rich prior knowledge compared to datasets that contain mainly basic entity types such as *PER/player*, *LOC/company*, etc. This can be seen in existing datasets where the label distribution of FET datasets is heavily skewed towards coarse-grained types. For example, the annotators of the OntoNotes [13] dataset labeled about half of the mentions as “other” because they could not find a suitable type for them. The */PERSON* label, on the other hand, covers more than 40% of the cases in the FIGER [16] dataset (see 4.1 for detailed analysis). Weischedel [17] annotated 2311 Wall Street Journal articles from Treebank-2 (LDC95T7) to construct the BBN dataset. However, there are just two levels of hierarchy in the BBN sample. Compared with previous FET datasets, the label distribution in CySets, our constructed dataset based on literatures in the cybersecurity domain, is more diverse and fine-grained, with a more distinct type hierarchy relationship.

In this paper, we propose a multimodal representation learning approach that includes a multi-level fusion encoder. As we show in our experimental results, existing models may face challenges in understanding cybersecurity mentions based purely on a single modal contextual representation. However, by mastering various modal features, the proposed multimodal representation learning approach can be applied to comprehend entities more fully. In summary, our contributions are summarized as follows:

1. Based on scientific literatures, we constructed the first manually annotated fine-grained cybersecurity entity typing dataset, with a stronger hierarchical structure and a more diverse label distribution.
2. A brand-new model called Cyst-MMET is proposed and evaluated, which can enrich entity mentions by multimodal knowledge representation learning and solve the data bottleneck and long-tail problems caused by fine-grained label sets.
3. A multi-level fusion encoder is designed and implemented. Through token-wise similarity calculation, fine-grained graphical matching is obtained to solve the semantic gap between different modalities and alleviate the problem of irrelevant visual noise, so as to learn a unified representation of different modalities. Experiments on all datasets show that our model achieves state-of-the-art (SOTA) or dominant performance (3%).

The rest of the paper is structured as follows. Section 2 reviews relevant existing models. Section 3 describes the proposed model in detail. Section 4 discusses the experimental setup and a detailed performance comparison with existing models, followed by concluding remarks in Section 5.

2 Related work

Because of its intelligence, automated recognition, and robust data analysis capabilities, artificial intelligence (AI) collaborates deeply with cybersecurity technologies and applications as it advances to the cognitive intelligence stage [18]. Entity typing is a typical task in AI and plays a key role in building knowledge hierarchy relationships in the security domain. The literatures [19–21] employ cybersecurity domain knowledge to analyze cybersecurity threats, such as building knowledge graphs from textual descriptions of cyber-attacks to better correlate data. However, most of these systems can only be seen as a black box to users. To improve our understanding of such systems, adversarial machine learning approaches can be used. The main features are detected by analyzing the extent of such changes, which helps in identifying the main reasons for misclassification. Sharma[46]presented approach has obtained satisfactory results that accurately explains the reasons for misclassifications.

Conducting Fine-grained entity typing (FET) task studies helps researchers to build a hierarchical body of knowledge and improve the effectiveness of learning. And it holds great promise in research areas of NLP tasks [22], such as link prediction and knowledge base construction. Nasiri et al. [50] proposed a novel Robust Graph Regularization Nonnegative Matrix Factorization for Attributed Networks (RGNMF-AN), which models not only the topology structure of networks but also their node attributes for direct link prediction. Different from FET, traditional NER is treated as a sequential labeling task [23, 44], limiting the number of entity classes. ASRNN [43] is a powerful tool for sequence labeling tasks that can effectively incorporate contextual information using attention mechanisms. The self-attention-based conditional random fields (CRF) latent variables model [45] for sequence labeling is a type of machine learning algorithm that is used to predict the labels of sequential data. However, the entity boundaries of the FET task are usually predefined, and it is generally treated as a hierarchical multi-label classification task. As a result, the former entity type identification approach clearly limits the breadth of information that can be extracted for entities. For example, the type of Security Algorithm can be subdivided into Encryption and Asymmetric subtypes. Researchers investigated entity typing tasks in various scenarios. Lin et al. [47] proposed a neural-encoded mention-hypergraph (NEMH) model to use hypergraph to model overlapping or nested structure mentions and use neural networks to extract features for hypergraph automatically which can effectively capture nested mention entities with unlimited length. Yao et al. [26] studied lexicon-level prediction, i.e., assigning a corresponding entity category to a noun phrase in the absence of context. Schütze [27] studied corpus-level prediction. Different from them, we focus on selecting the appropriate set of label types for mentions in a particular sentence.

Recent work has introduced fine-grained type ontologies to address the problem of large-scale fine-grained set of entity labels [24]. FIGER [16] defined 112 entity types based on Freebase (1 K), merging the categories with fewer entities. However, there are some problems with the FIGER [16] dataset, such as the extensive training set but only over 500 samples in the test set. As a result, OntoNotes [13] built a clearer hierarchy between entity categories based on this. Murty et al. [25] proposed TypeNet, a dataset with a deeper hierarchical structure that contains more than 1900 entity categories and filters some categories with fewer entities, based on WordNet (16 K). They do, however, focus on named entities, and data collection is challenging. In contrast, our ontology is based on external knowledge bases and literatures, contains nouns (and even

pronouns, as shown in Fig. 2b), goes beyond the traditional notions of named entities, and has a more diverse and fine-grained distribution of labels with domain expertise and credibility.

However, fine-grained label sets also lead to data bottlenecks and long-tail problems. In recent years, several approaches [28–30] have attempted to address this problem by introducing Zero-shot or Few-shot learning methods or data enhancement by removing label noise [31, 32]. Lv et al. [48] presented fine-grained Graph Auxiliary augmentation (GAU) model which trains the primary task together with an automatically created auxiliary task. And an auxiliary augmentation strategy was designed to enlarge the labeled set for the auxiliary task by utilizing the pseudo-labels of the primary task. This approach is used to solve the problem of degradation of classification models due to low training samples. Like the literature [33], we use an external knowledge base to introduce more external knowledge. In contrast, we consider knowledge from other modalities as the supplements, such as images in scientific literatures. Sun et al. [34] used chemical structure maps as auxiliary information for type identification of organics, while Azadifar et al. [49] proposed a novel graph theoretic-based gene selection method which was developed for cancer diagnosis. In this proposed method the optimal number of the final gene set was determined automatically. In line with previous studies [23, 34, 41], we utilize multimodal representations to alleviate the problem of unimodal information limitation, thus enriching entity mentions and providing context-sensitive fine-grained type labels.

3 Method

3.1 Task and data

Due to the domain's complexity and the research problem's cutting-edge nature, there is no standard dataset available for fine-grained cybersecurity entity typing. To this end, we collected recent cybersecurity research papers from the forthcoming paper platform WoS(Web of Science)¹ and annotated a new dataset, CySets, based on them.

We begin by obtaining image and text pairs from the literatures, with the text extracted using a string-matching method based on keywords and templates. The image and text sample pairs are created by matching the image numbers in the text with the image order in the literature. The text description primarily consists of a description and analysis of the information in the images' captions and texts below the images. The co-reference parsing system then selects mentions by extracting the longest noun phrase [35]. Finally, we provided image text pairs and entity mentions to five Ph.D. and four Master students in cybersecurity/information security majors and asked them to annotate the types of entities. To construct a hierarchical set of fine-grained entities, we require annotators to include a coarse-grained type (e.g., *Security Algorithm, Asset, Threat*) and at least one specific type (e.g., *Key Exchange, Resource, Malware*) for each mention. To improve consistency and accuracy, we had students from different grade levels annotate each selected scientific text twice, taking advantage of the annotators' prior knowledge. We defined F1 as a dataset evaluation metric to handle disagreement when calculating annotation alignment in two passes. We discarded the document as an "uncertain document" when the F1 score was

¹ <https://webofscience.com/>

Table 1 Statistics for the CySets Dataset

Model	Train	Dev	Test
Proportion	60%	20%	20%
Sentence	7745	2653	2961
Mention	8932	4007	5438
Level	Num.	Exam.	
Coarse-type	12	Asset	
Fine-grained	37	/Resource	
Ultra-fine	28	/Memory	
Other	20	//RSA	

less than a certain threshold. We chose a corresponding number of documents from the corpus at random to be re-annotated as a supplement. However, it is undeniable that cybersecurity knowledge is heterogeneous, and even annotators with a cybersecurity background may be unfamiliar with emerging or old relevant knowledge. To improve consistency even further, the final type set contains only 4/5 types annotated by annotators. We built the CySets development and test sets using the manual annotation approach described above.

In addition, we construct the training set of CySets by distant supervision according to Fig. 1. Common types from the cybersecurity literatures are first collected to construct a word list and then linked to Wikipedia² for expansion to improve the entity and type coverage of the knowledge base. Specifically, we construct a type tree to represent the entity types, and a complete set of fine-grained labeled types is the path from the root to the leaves in the tree structure, e.g. (*Asset*, */Resource*, */Memory*).³ We constructed the final data with more diverse labels and a sharper hierarchy than traditional datasets [13, 16]. The cleaned dataset data is shown in Table 1. However, CySets are not comprehensive, which makes the evaluation important (see Section 4 for details).

3.2 Model architecture

Understanding cybersecurity literatures is an interesting challenge from a NLP perspective. Our main idea to tackle this challenge is to perform multimodal representation learning and introduce an external knowledge base containing multimodal representations of security entities in cyberspace, such as algorithm principles, protocol flows, and natural language descriptions. Since different modalities of information are represented differently, we propose a Transformer-based multilevel fusion encoder to learn a unified multimodal representation to address the semantic gap between modalities. As shown in Fig. 3, the model can enrich entity mentions, enhance entity recognition, solve the data bottleneck and long-tail problems caused by fine-grained label sets, and alleviate the problem of irrelevant visual noise.

² https://en.wikipedia.org/wiki/Computer_security

³ In this paper, “/” denotes entity types at the Fine-grained level in CySets, “//” denotes the entity type at the Ultra-fine level in CySets

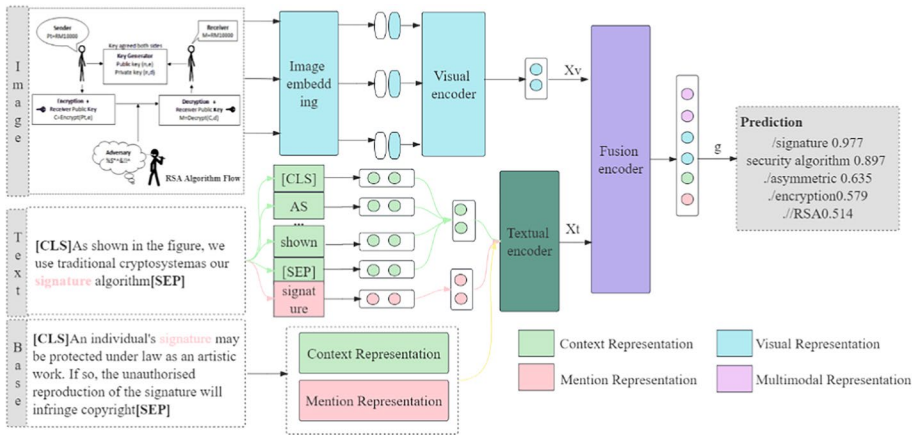


Fig. 3 Architecture of Cyst-MMET. This is a unified multimodal learning framework which includes three main components: text encoder, visual encoder, and fusion encoder. Different modal features are extracted separately and fed into the fusion encoder to obtain the unified representation vector

Context representation and mention representation In the FET task, contextual information is important. Determining entity subtypes without taking contextual information into account may introduce noise and lead to annotation ambiguity. Again, taking Fig. 2a as an example, we can quickly learn from the contextual semantics that [signature] refers to a cryptographic “signature algorithm” rather than a real-life “handwritten signature,” and thus we can determine that the subtype is probably RSA. Additionally, to address the OOV (Out of Vocabulary) problem, we use character-level embeddings in addition to the conventional word-level embedding. We employ the contextualized word representation ELMo [36] in place of earlier neural network models, which often use fixed-word embeddings, to conduct a fair comparison with some baseline models (e.g., LTR [11]). For encoding, we form the input instance as η , mark the entity with $[E_1]$, and end up with a sentence that looks like Eq. (1). In Eq. 1, a set refers to a collection of input tokens that are passed through the neural network model for processing. The CLS and SEP tokens are special tokens used in SciBERT [40] to indicate the beginning and ending of a sentence, respectively. The CLS token is added at the beginning of each input set to represent the classification task, while the SEP token is added at the ending of each set to separate it from the next set in the input sequence. The CLS and SEP tokens help SciBERT [40] to distinguish between different input sets and enable it to generate meaningful representations of the inputs that capture the context and meaning of the text.

$$\eta = \{[CLS], \eta_1, \dots, [E_1], m_1, \dots, m_k, [E_2], \dots, \eta_n, [SEP]\} \tag{1}$$

where m denotes the mention word and η_i denotes the context. Then, we feed η into SciBERT [40] and obtain the source hidden state $\varphi = \{\varphi_1 \dots \varphi_n\}$, where $\varphi \in \mathbb{R}^d$, d is the dimension of the hidden state. Finally, using the [CLS]-labeled hidden vectors as sentence embeddings, we obtain the representation vectors of context and mentions as V_{cm} .

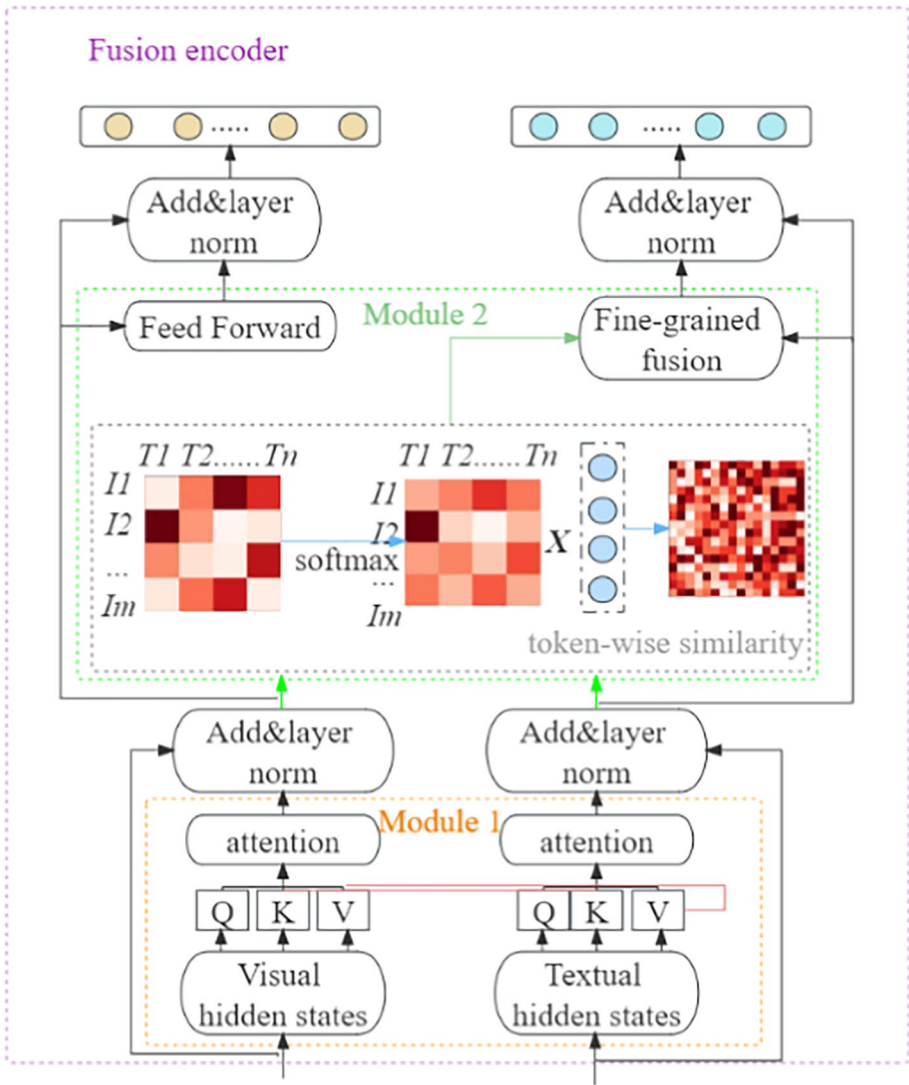


Fig. 4 Fusion encoder framework. It mainly consists of two parts: Module 1 is a coarse-grained interaction module, and Module 2 is a fine-grained fusion module

3.3 Multi-level fusion encoder

Inspired by CLIP [37] and Xu [23], we propose a multi-level fusion encoder to bridge the semantic gap caused by different modal representations. As shown in Fig. 4, we design a coarse-grained interaction module (Module 1) and a fine-grained fusion module (Module 2) to preemptively reduce the modal heterogeneity and mitigate the noise of irrelevant visual elements, respectively.

Specifically, we first redefine the calculation of multi-head attention at every layer to pre-reduce the modality heterogeneity, which is performed on the hybrid keys and values. We

reformat the equations as Eqs.(2) and (3),⁴ where $\lambda(x^v)$ denotes the scalar for the sum of normalized attention weights on the textual key and value vectors, as shown in Eq. (4). The overall sense of the interaction mechanism reduces the original visual attention probabilities and redistributes the remaining attention probability mass λ to focus on textual attention. CystMMET learns coarse-grained modal fusion to minimize modal heterogeneity beforehand by applying it to the attention calculation of hidden visual states and hidden textual states.

The w_q^t represents the query vector for text, which is obtained by multiplying the query matrix Q by the input vector x^t . The $head^{M^v}$ represents the output of the multi-head attention module M^v for visual vector, which takes as the input of the query vector w_q^t and the key-value pairs (w_k^t, w_v^t) in the sequence. $\lambda(x^v)$ represents a non-linear transformation applied to the input vector x^v , which is used to calculate the key and value vectors for the attention mechanism.

$$head^{M^t} = Attn(x^t w_q^t, x^t w_k^t, x^t w_v^t) \quad (2)$$

$$head^{M^v} = \text{softmax}(Q_v, [K_v]) \begin{bmatrix} V_v \\ V_t \end{bmatrix} = \underbrace{(1 - \lambda(x^v) Attn(Q_v, K_v, V_v))}_{\text{standard}} + \underbrace{\lambda(x^v) Attn(Q_v, K_t, V_t)}_{\text{cross-modal}} \quad (3)$$

$$\lambda(x^v) = \frac{\sum_i \exp(Q_v K_t^T)_i}{\sum_i \exp(Q_v K_t^T)_i + \sum_j \exp(Q_v K_v^T)_j} \quad (4)$$

The basic Transformer [38] configuration is used in this paper, where Q stands for the query mapping, (K, V) for the key-value pair, x^v for the visual feature vector, and x^t for the text feature vector, which are the output features corresponding to the interaction module. It is worth mentioning that the text feature vector consists of the text described in the literature and the text from the external knowledge base as Eq. (5), and each of them contains a contextual representation and a mention representation:

$$X_t = [V_{cm} : V_b] \quad (5)$$

where V_b stands for text feature vector of external knowledge base, X_t stands for feature vector of all text.

To mitigate the detrimental effect of noise, we calculate the similarity matrix of all text tokens as Eq. (6). We use a token-wise similarity calculation, as opposed to the CLIP [37] global similarity calculation, to capture fine-grained matching and localization of images with text at the token level. We then apply a SoftMax function to the similarity matrix of the i -th text token and use the average token aggregator of visual tokens in the image as Eq. (7), where Agg_i denotes the similarity-aware aggregated visual representation for i -th textual token.

$$S = x^t (x^v)^T \quad (6)$$

$$Agg_i(x^v) = \text{softmax}(S_i) x^v \quad (7)$$

⁴ The derivation process of the attention mechanism, only the key steps are written. Without loss of generality, the SoftMax scaling factor \sqrt{d} is ignored.

We also propose merging the similarity-aware aggregated visual hidden states into the text hidden states in the feed forward neural network layer (FFNN) and modifying the FFNN process’s computation to Eq. (8). W_3 in Eq. (8) represent the new added parameters for aggregated visual hidden states. The similarity matrix shows the nearest image patch for each text token. By inserting the aggregated visual representation based on similarity into the text-side FFNN algorithm to reduce the noise of unrelated entity images, a fine-grained alignment between image patch and text tokens is learned.

$$FFNN(x^t) = RELU(x^t W_1 + b_1 + Agg_i(x^v)W_3)W_2 + b_2 \tag{8}$$

3.4 Label predictions

We learn the type label embedding matrix with various granularities to evaluate our dataset in greater depth. The predicted type set has two cases: the label types with average maximum probability when both probabilities are less than 0.5 and the label types with the forecasted probability of more than 0.5. The threshold is set at 0.5.

$$y = \sigma(FFNN(W_t g))W_t \in \mathbb{R}^{n \times d} \tag{9}$$

$$g = [X_t : X_v : X_m] \tag{10}$$

In the above Eqs. (9) and (10), where n is the number of labels in the prediction space, d is the dimensionality of g , g is the final representation, and X_m denotes the fused multimodal representation. The label embedding matrix W_t consists of three granularity label sub-matrices of “Coarse-type, Fine-grained, Ultra-fine”, respectively, W_C, W_F, W_U .

3.5 Multi-task objective

Previous studies [9, 16] primarily employed special hinge-loss to improve noise or incomplete supervision robustness. Instead, we propose a multi-task objective that better reflects the training dataset’s fine-grained features. Instead of updating all labels for each example, we divide them into three granularities (Coarse-type, Fine-grained, Ultra-fine), with different granularities having different $\mathcal{L}_{granularity}$ training objectives. And we only update the labels in the granularity category with at least one label. Specifically, the training goal is to reduce \mathcal{L} , as shown in Eqs. (11) and (12). In Eq. 11, the loss function is defined as the cross-entropy loss between the predicted class probabilities \hat{y}_i and the true class labels y_i for each input example i in the training set. In Eq. 12, the loss function is defined as multi-task objective training for three granularities.

$$\mathcal{L} = - \sum \log(y_i) + (1 - t_i) \cdot \log(1 - y_i) \tag{11}$$

$$\mathcal{L}_a = \mathcal{L}_C \cdot \Psi_C(t) + \mathcal{L}_F \cdot \Psi_F(t) + \mathcal{L}_U \cdot \Psi_U(t) \tag{12}$$

The function $\Psi_{granularity}(t)$ checks whether a given granularity’s target vector t contains the entity type. Here, t is the target vector for each granularity.

The overall process for the FET task in our model is shown in Algorithm 1. The inputs mainly include Multiple datasets \mathcal{D} such as CySets, OntoNotes, BBN, FIGER; the batch

size b , the temperature parameter τ , the learning rate α , the score function $\psi(\cdot)$. Our goal is to predict the type of entity based on the input entity mentions and context. Our model goes through three main processes. Firstly, text and visual features are extracted, where text features are extracted via SciBERT and visual features are extracted via a visual coder. The extracted features are then fused, which are achieved by fusing the coarse-grained alignment module and the fine-grained fusion module of the encoder. Finally the possible entity types are predicted from the obtained uniform representation.

Input: \mathcal{D} , Multiple datasets // *CySets, OntoNotes, BBN, FIGER*
 The batch size b , the temperature parameter τ ;
 The learning rate α , The score function $\psi(\cdot)$;
Output: The label prediction probability $y(\cdot)$

1 Feature Extraction Module:
 2 Procedure Textual Extraction // *Using in literature and external base*
 3 **for** $n = 1$ to N **do**
 4 $F_c \leftarrow \text{SciBERT}(f_c)$
 5 **return** (F_c)
 6 $F_m \leftarrow \text{SciBERT}(f_m)$ // *Extract mention feature with SciBERT*
 7 **return** (F_m)
 8 $X_t \leftarrow \text{concatenate}[F_c, F_m]$
 9 Procedure Visual Extraction
 10 **for** $n = 1$ to N **do**
 11 $X_v \leftarrow \text{Resnet}(f_v)$ // *Extract visual feature with Resnet*
 12 **return** (X_v)
 13 **Feature Fusion Module:** // *The output of text encoder and visual encoder as input*
 14 **for** $n = 1$ to N **do** // *N is the attention head*
 15 **for** $X_t \in \chi$ **do** // χ and v are vector spaces respectively
 16 **for** $X_v \in v$ **do** // *See details in Eq.(2)-Eq.(8)*
 17 $\text{head}^{M_t} = \text{Attn}(x^t w_q^t, x^t w_k^t, x^t w_v^t)$ // *Module1*
 18 $\text{head}^{M_v} = (1 - \lambda(x^v)) \text{Attn}(Q_v, K_v, Q_v) + \lambda(x^v) \text{Attn}(Q_v, K_t, Q_t)$
 19 $S = x^t (X^v)^v$ // *Module2*
 20 $\text{Agg}_i = \text{softmax}(S_i) x^v$
 21 **return** (X_m)
 22 $g \leftarrow \text{concatenate}[X_t, X_v, X_m]$
 23 **Prediction Module :**
 24 **for** $n = 1$ to N **do**
 25 $y = \text{sigmoid}(FFNN(W_t g))$
 26 **if** $y_t > 0.5$ // *Predict every type t for which $y_t > 0.5$*
 27 $\hat{y} = y$
 28 **else** $\hat{y} = \text{argmax}(y)$
 29 **return** \hat{y}
 30 From Eq.(11) get \mathcal{L} as the total loss function for our task.
 31 Back propagation and update parameters by Adam optimizer.
 32 **Repeat:** Repeated training to get the best predicted value.

Algorithm 1 The overall process for the FET task in our model

For the text, we use pre-trained SciBERT to encode the context and mentions to obtain the feature vectors of the two respectively, which are connected to form the text feature vector. The visual coding Resnet is used to encode the image regions. The same settings are used for the external knowledge base and the text in the literature. The multilevel fusion encoder takes the output of the text encoder and the visual coder as input, and the modal heterogeneity is pre-reduced by Module1, and fine-grained fusion is done by Module2 (see details in 3.3). Also, we set the threshold 0.5 to predict the label set.

4 Experiments

4.1 Settings

Datasets In addition to the experimental analysis of the dataset CySets constructed in this paper, we chose three standard fine-grained typing datasets, OntoNotes [13], FIGER [16], and BBN [17], to evaluate the performance of the proposed model Cyst-MMET.⁵The original version in OntoNotes[13]contains 25 K/2 K/9 K training/development/test data, 89 categories, and 2.7 labels per sample on average. The classification system is divided into three layers: person (first), artist (second), and actor (third). We used Shimaoka's [9] training, development, and test sets. The FIGER dataset, like OntoNotes [13], was created by merging categories with fewer entities from Freebase, and it includes 2.7 million automatically labeled training instances from Wikipedia and 434 manually labeled sentences from news reports. We partitioned the dataset similarly to Shimaoka [9]. The Wall Street Journal text corpus of one million words (LDC95T7) was annotated with a two-level hierarchy using the BBN [17] dataset to construct a corpus of entity types with a core reference of BBN nouns. We followed Ren and Zhang's methods [38, 39] for partitioning the datasets.

Baselines

- 1) ATTENTIVE [9]: Entity mentions and contexts are modeled separately to obtain the mention representation and the context representation. Where the entity mentions representation is directly averaged, the contextual representation uses a Bi-LSTM and fixed attention mechanism, and the spliced features are sent to the MLP(Multilayer Perceptron) for category prediction. It can be seen that this approach does not consider the problem of data noise caused by distant supervision.
- 2) BERT-CRF [39]: It is a multi-layer bidirectional Transformer encoder with SoftMax decoder. BERT-CRF is based on BERT with CRF decoder instead of SoftMax decoder.
- 3) VisualBERT [41]: VisualBERT consists of a bunch of Transformer layers that implicitly align the elements of the input text with the regions in the associated input image in a self-attention manner, allowing unsupervised association of language elements with image regions.
- 4) SciBERT [40]: SciBERT is a BERT model trained on 1.14 million papers from Semantic Scholar, of which 136,800 are from the computer science domain.

⁵ <https://github.com/INK-USC/PLE/blob/master/Data/README.md>

- 5) FGCEt [34]: Based on SciBERT, chemical structure graphs, in addition to context-based representation embedding, are used as auxiliary features to recognize entity types.
- 6) LTR [10]: It uses more robust pre-trained language models such as ELMo and BERT for entity mentions and contextual representations. The core of the model is a hybrid classifier that exploits the type interdependence of potential type representations. Instead of predicting each type independently, it predicts low-dimensional vectors encoding potential type features, and the model reconstructs sparse high-dimensional type vectors from such potential representations.
- 7) MAF [23]: It proposes a generic multimodal matching and alignment framework that reduces the impact of mismatched text-image pairs and makes the representation between two modalities more consistent. The modal matching and alignment modules are based on self-supervised learning and do not require additional data annotation. The different module feature vectors are connected to achieve modal fusion.
- 8) AFET [38]: It first extracts the features of mentions and then divides the training dataset into clean and noisy sets. The category information of entities in the clean set is relatively single, corresponding to only one category path, while the entities in the noisy set correspond to multiple category paths. This kind of data is the one containing noise. Mapping entity mentions and types into the same semantic space is convenient for doing calculations later. The training objective is to learn these two mapping matrices. After obtaining these two matrices, the categories can be predicted for the mentions in the test set.
- 9) MLR [12]: This method includes ontology structure in both training and prediction processes. In the training process, a new multilevel learning ranking loss is used to compare positive types with negative types based on a type tree. During prediction, a coarse-grained to fine-grained decoder which restricts the optional candidate objects at each level of the ontology based on the already predicted parent nodes (types).
- 10) NDP-PTC [32]: NDP is a new FET model that models the relationship between hierarchical types and noise. NDP-PTC is a progressive training method for training models that remove noise types from the training set.

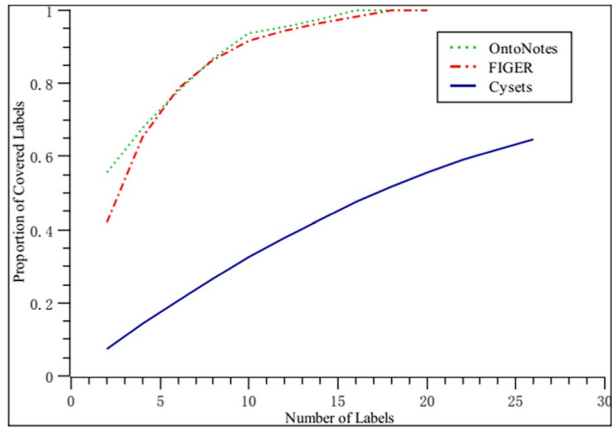
Implementation We train our model on a single GTX 3090 GPU with fp16. Specifically, we set the word embedding dimension size to 512, and the max tokens is set to 1536 and the attention head is 4. We use SciBERT-Scivocab (uncased)⁶ as our text encoder, with the learning rate set to 5e-5, other parameters set to 1e-3, and batch size set to 32. We also employ the Adam [42] optimizer, with an epsilon of 1e-6 and a warmup rate of 0.08. To reduce overfitting, we set the dropout to 0.2. Furthermore, the original-5.5b ELMo⁷ model is pre-trained and its weights are frozen during training to ensure a fair comparison with the baseline model. And we also use GraphPad prism and Piktochart as mapping tools.

Similar to previous work [9–12], we use precision, recall and F1 to evaluate the type label set at each granularity, and other performance through strict accuracy (Acc), macro-average F-score (Macro F1) and micro-average F-score (Micro F1). For the i -th instance, let the set of the true types be T_i , and the set of the predicted types be \hat{T}_i . The strict accuracy is the ratio of instances where $T_i = \hat{T}_i$. Macro F1 is the average of all F1 scores between T_i and

⁶ <https://github.com/allenai/scibert>

⁷ <https://allennlp.org/elmo>

Fig. 5 Distribution of labels for different evaluation datasets. In the OntoNotes and FIGER datasets, 4–7 types of labels alone cover more than 80% of the samples. In CySets, the first 26 tags cover less than 80% of the data



\hat{T}_i for all instances, whereas micro F1 counts total true positives, false negatives and false positives globally.

For Micro F1:

$$MicroF1 = 2 \frac{Recall * Precision}{Recall + Precision} \tag{13}$$

$$Precision = \frac{\sum_{i=1}^N |T_i \cap \hat{T}_i|}{\sum_{i=1}^N |\hat{T}_i|} \tag{14}$$

$$Recall = \frac{\sum_{i=1}^N |T_i \cap \hat{T}_i|}{\sum_{i=1}^N |T_i|} \tag{15}$$

For Macro F1:

$$Precision = \frac{1}{N} \sum_{i=1}^N \frac{|T_i \cap \hat{T}_i|}{|\hat{T}_i|} \tag{16}$$

$$Recall = \frac{1}{N} \sum_{i=1}^N \frac{|T_i \cap \hat{T}_i|}{|T_i|} \tag{17}$$

4.2 Data analysis

As shown in Table 1, we construct our dataset CySets by distant supervision and manual annotation using scientific literatures in the domain of cybersecurity as a data source. The entity types are subdivided into several granularities for the analysis, and we concentrate on three granularities: coarse-type, fine-grained, and ultra-fine, with a combined total of

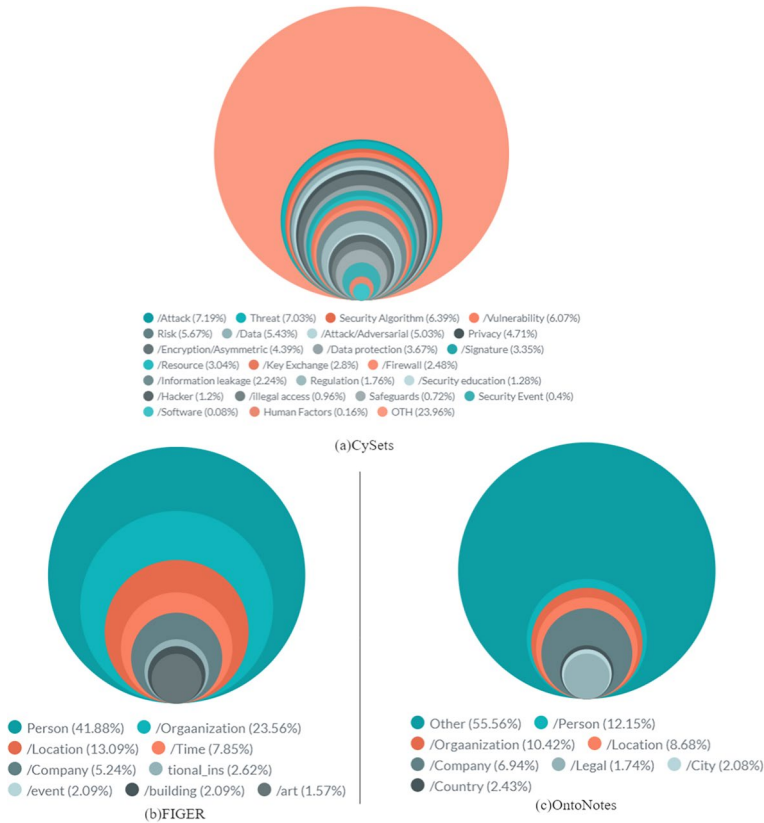


Fig. 6 Visualization results of different datasets. It is worth noting that “OTH” in CySets does not refer to a separate category, but rather represents a collection of entity labels other than the entity types listed in the figure

77 labels. Google drive⁸ has the entire collection of labels, where various colors stand for various granularities.

- **12 Coarse-type types:** Asset, Threat, Risk, Security Algorithm, Safeguards, Privacy, Exposure, Human Factors, Privacy, Regulation, Threat Agent, Security Event
- **37 Fine-grained types:** As first-level subtypes of Coarse-type types, e.g., /Resource, /Attack.
- **28 Ultra-fine types:** As secondary subtypes of Coarse-type, e.g., /Memory, /Syntactic attacks.

We compare CySets with previous typical FET datasets, and the results show a more diverse and fine-grained label distribution in our dataset. Figure 5 shows the percentage of labels covered by the top N labels in each dataset.

⁸ <https://drive.google.com/file/d/1mNM0UEt7D-e9hsMEVRQzxJ277Dt5GB7v/view>

Table 2 Performance of the FET model on the CySets dataset

Modality	Model	CySets		
		Accuracy	Micro-F1	Macro-F1
Text	ATTENTIVE [9]	14.12	39.25	39.90
	BERT [40]	16.37	41.26	42.39
	BERT-CRF [39]	18.44	43.93	45.08
	Cyst-MMET-T	18.75	44.11	45.32
Text+Image	VisualBERT [41]	18.81	42.88	45.03
	SciBERT [40]	18.77	42.81	43.89
	FGCET [34]	18.84	42.90	44.75
	LTR [10]	19.71	43.63	45.27
	MAF [23]	19.06	42.67	44.59
	Cyst-MMET	22.73	47.30	51.09

It is obvious that the curve of CySets grows more slowly and converges more slowly. In contrast, the OntoNotes and FIGER datasets, as the control group, converge extremely fast, and the distribution of labels is highly skewed towards the first few labels. Specifically, the FIGER dataset covers 80% of the samples in just 7 entity types, while OntoNotes is better, requiring only 4 types. We speculate that this may be due to its predefined ontology that limits the number of entity categories. In our CySets, on the other hand, the first 26 labels cover only about 76% of the data.

We conducted additional research using visualization to highlight the fine-grained hierarchy and diversity of CySets entity type labels and to more visually show the differences across datasets in Fig. 6. We used the statistical mapping tool PICTOCHAT to analyse the labels in the different datasets. For the Cysets dataset that we constructed, we selected 23 most frequent tags as representatives, and the rest of the tags were grouped into the “OTH” category, and their proportion of the total number of tags was calculated to visualize the tag coverage. For the FIGER and OntoNotes datasets, we used data from the literature [13, 16], respectively. The label distributions of the previous FET datasets (FIGER and OntoNotes) are clearly skewed towards coarse-grained types, with a clear long-tail distribution. In the FIGER dataset, for example, the coarse-grained type “Person” covers more than 40% of the entity types. It is worth noting that roughly half of the entity mentions in the OntoNotes dataset are labeled as “Other,” owing to the fact that many mentions cannot be found to correspond to the ontology and must be “passively” grouped together. In contrast, our dataset has a more distinct hierarchy and more extensive labeling.

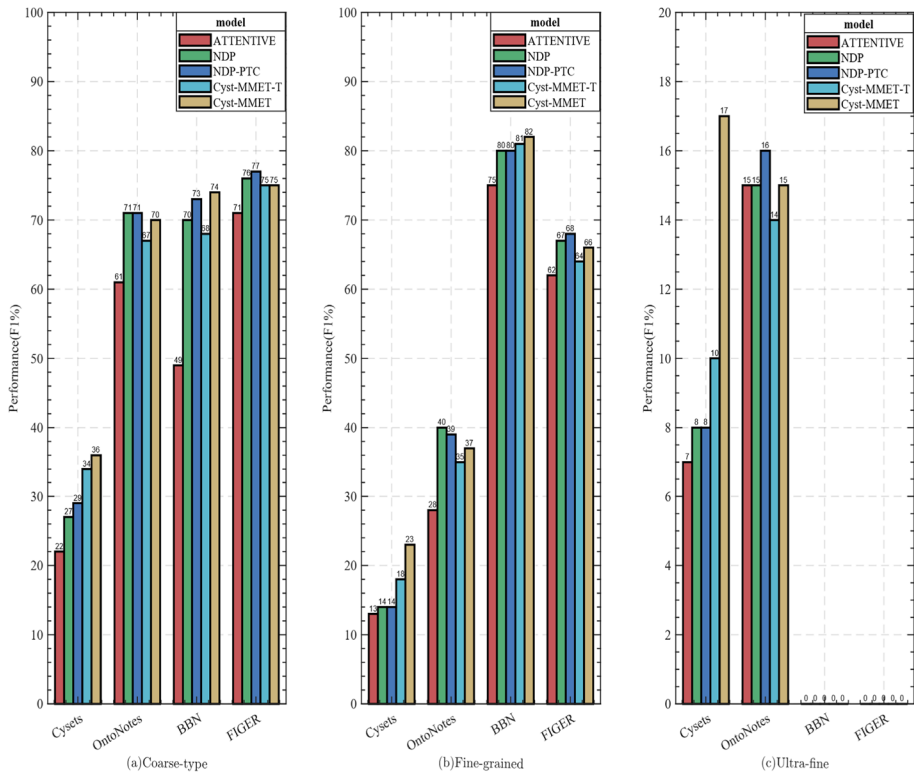
4.3 Comparative experiment

To be fair, we divide the proposed model into two sets and test them with the baseline models equally. Table 2 displays the experimental findings of the comparison with the baseline models on the CySets dataset. Cyst-MMET-T is the text-only model without the addition of multimodal information, and Cyst-MMET is the model with the whole modules.

Table 2 demonstrates that the suggested model in this paper still outperforms baseline multimodal baseline models in the Micro-F1 and Macro-F1 measures, even after removing the visual encoding component from both the standard BERT and BERT+CRF models on the CySets dataset. We hypothesize that there are two main causes for this. The first is that

Table 3 Results of our model for development sets with different granularity

Granularity	Precision	Recall	F1
Coarse-type	35.7	37.1	36.4
Fine-grained	23.6	21.9	22.7
Ultra-fine	26.3	6.6	10.6

**Fig. 7** Performance at different granularities on three test datasets

the model suggested in this research uses data from external knowledge bases as the supplements, and the candidate set coverage is broader to ease the typing task. Second, Cyst-MMET-T considers both word embedding and character embedding, thus enabling finer-grained potential information to be obtained. In contrast to our multilevel fusion encoder, MAF [7] simply connects different modal features during fusion, ignoring the intrinsic semantics between modalities. This is later confirmed by the SOTA achieved by our model on CySets. In conclusion, the scope and accuracy of entity classification candidate sets are improved by multimodal augmented text representation, alleviating the data bottleneck and long-tail problems.

Table 3 illustrates the performance decomposition of the entity hierarchy on the development set at various granularities. As the entity hierarchy deepens and the granularity refines, the model performance deteriorates. As previously demonstrated in the fine-grained NER literature [13, 38], fine-grained labels are more difficult to predict

Table 4 Performance comparison of different models on three standard datasets

Model	OntoNotes			BBN			FIGER		
	Acc	Mi	Ma	Acc	Mi	Ma	Acc	Mi	Ma
AFET [38]	51.1	64.7	71.1	67.0	73.5	72.7	53.3	66.4	69.3
ATTENTIVE [9]	51.7	64.9	71.0	–	–	–	59.7	75.4	79.0
LTR [11]	63.8*	77.3*	82.9*	55.9	79.3	78.1	62.9	79.8	83.0
MLR [12]	58.7	68.1	73.0	75.2	79.7	80.5	65.5	78.1	80.5
NDP [32]	59.2	66.3	72.6	77.1	82.1	81.3	68.6	78.8	82.1
NDP-PTC [32]	59.6	66.9	73.2	77.9	81.9	82.3	70.0	79.5	82.6
Cyst-MMET	59.3	69.5	71.8	78.1	83.0	80.4	69.8	80.4	81.9

–: Not run on the specific dataset.

*:Not strictly comparable due to non-standard, much larger training set

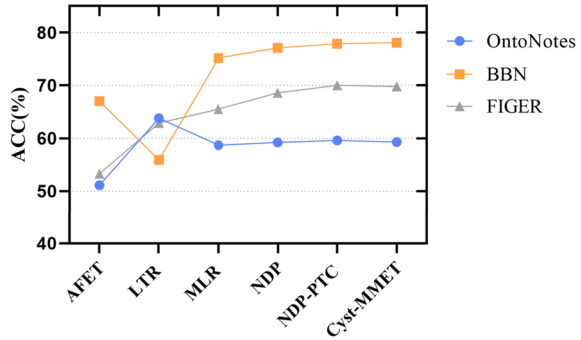
than coarse-grained labels, and this problem is exacerbated when dealing with ultra-fine types. As detailed in Section 4.5, the lower performance is due in part to noun/pronoun mentions (e.g., “he”), as well as synonyms, annotation ambiguities, and so on.

We also show the performance at different granularities on the three standard fine-grained typing datasets in Fig. 7. The F1 score is used as an evaluation metric. Specifically, for Fine-grained granularity, performance is evaluated for all datasets. Also, since BBN does not have any noise in the test dataset, the entity does not have the label of Ultra-fine granularity. However, for the FIGER and OntoNotes datasets, an entity can be assigned to different levels at the same time. Moreover, our model achieves SOTA performance on CySets and BBN datasets, both of which are based on scientific literature annotations, confirming the effectiveness of the model on literature-based datasets. Moreover, comparing the experimental results of Cyst-MMET and Cyst-MMET-T in CySets, we can find that the growth of F1 score increases gradually with the finer granularity of the dataset. This proves that our multimodal model has a more pronounced enhancement at finer granularity and is more effective for expanding finer granularity type sets.

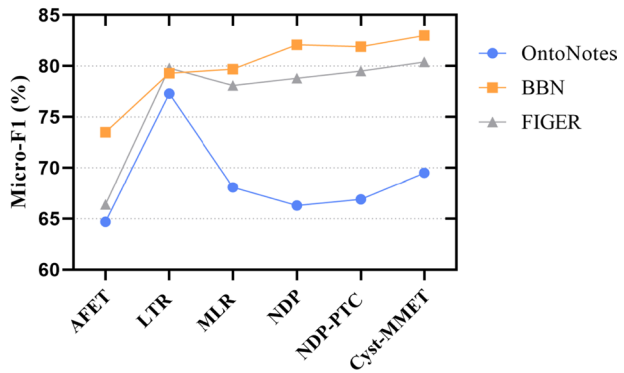
By comparing Cyst-MMET with other models on three standard datasets, Table 4 provides the results. By utilizing the ACC, Macro-F1, and Micro-F1 evaluation metrics, we were able to confirm the model’s validity. The best results are displayed in bold for each dataset and evaluation metric. Special situations are described at the bottom of the table, and the performances of the baseline models are taken from the literatures. We will then present the visualization in Fig. 8 and analyze it further.

Our method, in particular, performs SOTA in Micro-F1 scores (see Fig. 8b), the most advanced or dominant performance ($\pm 3\%$) in Accuracy (see Fig. 8a), and a micro difference in Macro-F1 scores ($<2\%$). We hypothesize that this is due to the addition of a noise cleaning module (NDP-PTC [10]), which improves the “purity” of the labels, and this is supported by the significant advantage achieved by Cyst-MMET over NDP (without the noise cleaning module). Furthermore, when compared to previous models, such as MLR [12], SOTA performance is obtained across all datasets and metrics. On the FIGER dataset, in particular, our method significantly improves the accuracy score (+4.3%), indicating that our model is capable of producing a more accurate type set.

Fig. 8 Evaluation results of different models on OntoNotes, BBN, and FIGER datasets. **a** accuracy, **b** Micro-F1

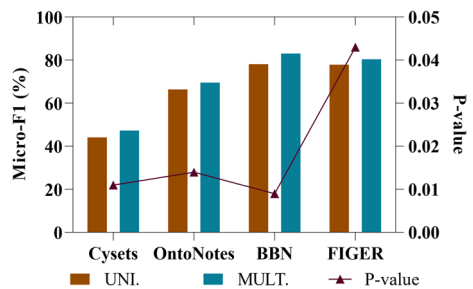


(a) Accuracy in three different datasets



(b) Micro-F1 in three different datasets

Fig. 9 Performance of unimodal and multimodal on different datasets. UNI. denotes the model Cyst-MMET-T and MULT. denotes the multimodal model Cyst-MMET



4.4 Modal enhancement

As shown in Fig. 9, we verified the effectiveness of the multimodal model on CySets and three standard datasets, respectively. Among them, the improvement is most obvious on the BBN dataset, reaching nearly 5%, and about 3% on the other datasets, which fully demonstrates the effectiveness of multimodality. We conjecture that the reason for the best results on BBN is that BBN does not have any noise labels in the test dataset,

Table 5 Ablation results on the CySets

Model	CySets		
	Accuracy	Micro-F1	Macro-F1
Full MODEL	22.73	47.30	51.09
w/o visual	18.75	44.11	45.32
w/o description	19.53	45.06	47.55
w/o fusion encoder	21.39	46.24	49.17
w/o M1	21.55	46.82	49.96
w/o M2	21.47	46.45	49.33

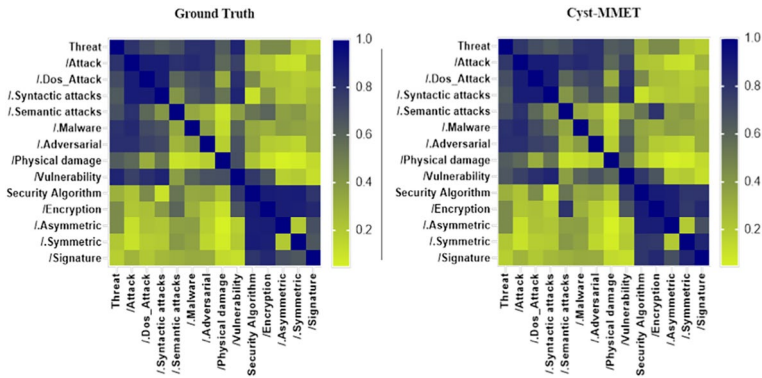


Fig. 10 Heat map of model prediction results and co-occurrence matrix of Ground Truth. Cyst-MMET learns co-occurrence matrix similar to Ground Truth

which is extremely friendly to our model without the noise reduction module. Thus, it is undeniable that our multimodal component is effective for the FET task.

In addition, we also investigated the p-values of the Micro-F1 significance t-test between UNI (i.e., Cyst-MMET-T) and MULT (i.e., Cyst-MMET) on the four datasets of 0.011, 0.014, 0.009, and 0.043, respectively. Thus, we can conclude that Cyst-MMET significantly outperforms Cyst-MMET-T for almost all metrics on all four datasets. This result confirms the significant advantage of our Cyst-MMET model.

4.5 Ablation experiment

We conducted ablation experiments on the proposed model, as shown in Table 5, and the results of the ablation experiments show that the proposed model in this paper is sensitive to each module. Specifically, the impact of removing the visual modal information is greater than that of removing the text modality. The impact of Module2 in the multilevel fusion encoder is greater than that of Module1, because Module2 mitigates the noise of irrelevant entity images and achieves finer-grained cross-modal interaction compared to Module1. It is worth mentioning that w/o visual and w/o description remove the information of visual and text modality, respectively, and naturally there is no fusion module behind, so the experimental effect is the worst. This demonstrates the effectiveness of image enhancement for textual representation and multimodal learning.

Fig. 11 An example of prediction results, where entity mentions are marked with a green underline, correct predictions are shown in green font, and missing labels are indicated in red font. Also, black font indicates labels that appear in the prediction results but are not annotated

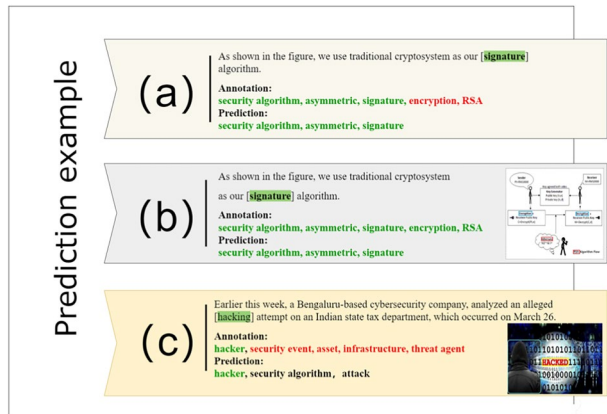
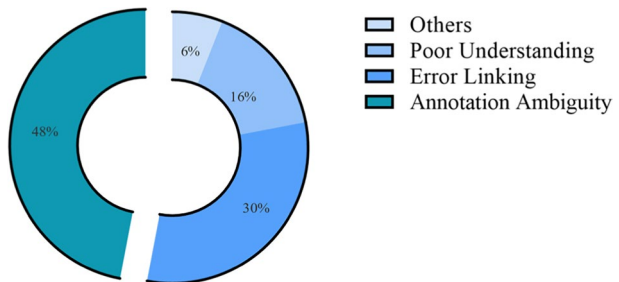


Fig. 12 Analysis of possible causes of errors



4.6 Case study

To visualize the relationship between the model prediction results and Ground Truth, we show the label co-occurrence matrix between the model prediction results and Ground Truth in Fig. 10. The greater the correlation index, the darker the color represents, indicating that the two labels are more similar. The diagonal line of the matrix represents two identical label correlations, as well as the maximum similarity of 1 between different labels. Furthermore, the closer the color in the two matrices, the smaller the difference between the prediction results and the fundamental. The difference between the two images is that the left image demonstrates similarity between real labels based on manual annotation, which leads to better separation and clustering of entity types based on their similarities. Whereas the figure on the right shows the relationship between the predicted labels of the multimodal model we have constructed. We aim to investigate the discrepancy between the predicted labels and the true labels with two graphs. We can see that Cyst-MMET can largely reproduce the label annotation results in Ground Truth, correlation can learn entity type labels accurately.

Furthermore, we manually examined 50 examples from the development set, three of which are depicted in Fig. 11. Overall, the model can produce an accurate set of entity type labels. We use Figs. 11a and 11b as a comparison group, where the former uses only textual information while the latter adds images matching the textual information. Figure 11c depicts an example of an annotation error. In (a), we use only textual information for the prediction, which is incomplete compared to the annotation (missing *encryption* and *RSA*).

However, when we introduce images as auxiliary information, this problem is solved and the prediction results are in full agreement with the annotations, thus demonstrating the enhanced effect of multimodality.

Despite our efforts to annotate a comprehensive set of fine-grained types, many potentially correct labels have been missed for a variety of reasons. For example, in Fig. 11c, “*attack*” makes sense, but is considered incorrect (Poor Understanding) and so does not appear in the annotation results. The dataset also contains synonym errors (not shown in Fig. 11), such as “Three years ago, [*the explosions*] occurred because of internal staff operational errors. “ This should be annotated as “*security event*” instead of “*attack*”, which may be the correct type but is not supported by the context (Annotation Ambiguity). In addition, there is an error called “Error Linking.” Although distant supervision is used to link context-independent entities to external knowledge bases for annotation, the linking error can be caused by the real-time and advanced nature of entities in the literatures, and the possible causes of the error are shown in Fig. 12. In conclusion, compared to other error types, annotation ambiguity due to insufficient understanding is the most common cause of errors (48%), which may be limited by the annotator’s knowledge base and lack of ability to separate some synonyms. In addition, since researchers may propose some new concepts, algorithms in the cybersecurity literatures, the external knowledge base don’t update this knowledge, leading to error linking, which is the second main cause of error (30%).

5 Conclusion and future work

In this paper, we provide an in-depth study of FET in the cybersecurity domain. A new dataset, CySets, is created based on the literatures to facilitate the study of new tasks in cybersecurity domain. Our dataset has sharper hierarchical relationships and more diverse fine-grained labels than the traditional FET datasets. We also propose a multimodal representation learning model, Cyst-MMET, that includes a multi-level fusion encoder to effectively integrate multimodal data and enhance the model’s understanding of cybersecurity domain knowledge. Experimental results show that our model can effectively utilize multimodal information to enrich the representation of cybersecurity entities, provide context-sensitive fine-grained type labels, and solve the data bottleneck and long-tail problems caused by fine-grained label sets. The fusion encoder can bridge the semantic gap between different modalities and alleviate the problem of irrelevant visual noise. In addition, the multimodal entity representation learning approach proposed in this paper is general enough to be used for entity typing tasks in other domains. Admittedly, an existing challenge in this paper is that many cybersecurity entities cannot be linked to external knowledge bases because they simply do not contain that particular entity, which is especially evident for new concepts in the cybersecurity literatures. Therefore, in future work, we consider introducing better entity linking algorithms to enhance the matching of entity mentions with external knowledge bases. In addition, the noise problem in all FET tasks still deserves our attention. We only consider which labels are mistakenly considered as “noise” to be added to the candidate type set, so in the future we intend to introduce a noise cleaning module into the model to complete the noise reduction/denoising of the dataset more completely.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant No. 61862063, 61502413, 61262025; the Science Foundation of Young and Middle-aged Academic

and Technical Leaders of Yunnan under Grant No. 202205 AC160040; the Science Foundation of Yunnan Jinzhi Expert Workstation under Grant No. 202205AF150006; Major Project of Yunnan Natural Science Foundation under Grant No. 202202AE090066; Science and Technology Project of Yunnan Power Grid Co., Ltd. under Grant No. YNKJXM20222254; the Science Foundation of “Knowledge-driven intelligent software engineering innovation team”.

Authors’ contributions In this paper, we constructed a fine-grained entity typing dataset based on multi-modal information from the cybersecurity literatures and design a multimodal representation learning model based on it. Baolei Wang completed the model design and experimental analysis of this paper and wrote the core chapters of the paper. Xuan Zhang is the corresponding author of this paper, providing hypothetical opinions for this article. Gao Chen completed the second chapter of this article, and Jishu Wang and Linyu Li completed the Figs. 9, 10 and 11 in the fourth part. Qing Duan provided grammatical help for the writing of this article. All authors composed the rest of the manuscript and reviewed the whole manuscript.

Funding This work was supported by the National Natural Science Foundation of China under Grant No. 61862063, 61502413, 61262025; the Science Foundation of Young and Middle-aged Academic and Technical Leaders of Yunnan under Grant No. 202205 AC160040; the Science Foundation of Yunnan Jinzhi Expert Workstation under Grant No. 202205AF150006; Major Project of Yunnan Natural Science Foundation under Grant No. 202202AE090066; Science and Technology Project of Yunnan Power Grid Co., Ltd. under Grant No. YNKJXM20222254; the Science Foundation of “Knowledge-driven intelligent software engineering innovation team”.

Data availability All data generated or analyzed during this study are included in this published article.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval Not applicable.

References

1. Bridges R A, Jones C L, MD Iannacone, et al. (2013) Automatic labeling for entity extraction in cyber security[J]. *Comput Sci*
2. Joshi A, Lal R, Finin T, Joshi A (2013) “Extracting cybersecurity related linked data from text,” in *Proceedings of the 7th IEEE International Conference on Semantic Computing*. IEEE Comput Soc Press
3. Huang S, Sha Y, Li R (2022) A Chinese named entity recognition method for small-scale dataset based on lexicon and unlabeled data[J]. *Multimed Tools Appl*:1–22
4. Choi E, Levy O, Choi Y, Zettlemoyer L. (2018) Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, volume 1: long papers, pages 87–96*. Association for Computational Linguistics
5. Del Corro L, Abujabal A, Gemulla R, Weikum G. (2015) FINET: context-aware fine-grained named entity typing. In *Proceedings of the 2015 Conference on empirical methods in natural language processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015, pages 868–878*. The Association for Computational Linguistics
6. Zhang S, Balog K, Callan J (2020) “Generating categories for sets of entities,” in *Proc ACM Conf Inf Knowl Manage*, pp. 1833–1842
7. Onoe Y, Durrett G (2020) Fine-grained entity typing for domain independent entity linking. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, pages 8576–8583*. AAAI Press. LNCS Homepage, <http://www.springer.com/lncs>, last accessed 2016/11/21
8. Yavuz S, Gur I, Su Y, Srivatsa M, Yan X. (2016) Improving semantic parsing via answer type inference. In *Proceedings of the 2016 Conference on empirical methods in natural language processing*,

- EMNLP 2016, Austin, Texas, USA, November 1–4, 2016, pages 149–159. The Association for Computational Linguistics
9. Shimaoka S, Stenetorp P, Inui K, Riedel S (2017) Neural architectures for fine-grained entity type classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3–7, 2017, volume 1: long papers, pages 1271–1280. Association for Computational Linguistics
 10. Lin Y, Ji H (2019) An attentive fine-grained entity typing model with latent type representation[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 6197–6202
 11. Chen T, Chen Y, Van Durme B (2020) Hierarchical Entity Typing via Multi-level Learning to Rank. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8465–8475, Online. Association for Computational Linguistics
 12. Xu P, Barbosa D (2018) “Neural fine-grained entity type classification with hierarchy-aware loss,” in Proc. Conf. North Amer Chapter Assoc Comput Linguistics, pp. 16–25
 13. Gillick D, Lasic N, Ganchev K, Kirchner J, Huynh D (2014) Context dependent fine-grained entity type tagging. CoRR abs/1412.1820:1–9
 14. Raiman J R, Raiman O M (2018) Deep type: multilingual entity linking by neural type system evolution[C]. Thirty-Second AAAI Conference on Artificial Intelligence
 15. Sun C, Li W, Xiao J, et al. (2021) Fine-grained chemical entity typing with multimodal knowledge representation[J]
 16. Ling X, Weld DS (2012) Fine-grained entity recognition. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)
 17. Weischedel R, Brunstein A (2005) BBN pronoun coreference and entity type corpus[J]. Linguistic Data Consortium, Philadelphia, p 112
 18. Fang B, Shi J, Wang Z et al (2021) Security threats and countermeasures of artificial intelligence-enabled cyber attacks [J]. China Eng Sci 23(3):7
 19. Pingle A, Pillai A, Mittal S, et al. (2020) Relet: relation extraction using deep learning approaches for cybersecurity knowledge graph improvement[C]// 2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE
 20. Kang Y, Zhong J, Li R, et al. (2021) Classification method for network security data based on multi-featured extraction[J]. Int J Artif Intell Tools
 21. Shen G, Wang W, Mu Q et al (2020) Data-driven cybersecurity knowledge graph construction for industrial control system security[J]. Wirel Commun Mob Comput 2020(6):1–13
 22. Raiman J, Raiman O (2018) Deeptype: multilingual entity linking by neural type system evolution. In Association for the Advancement of Artificial Intelligence
 23. Xu B, Huang S, Sha C et al (2022) MAF: a general matching and alignment framework for multimodal named entity recognition[C]//proceedings of the fifteenth ACM. Int Conf Web Search Data Min:1215–1223
 24. Rabinovich M, Klein D (2017) Fine-grained entity typing with high-multiplicity assignments. In proceedings of Association for Computational Linguistics (ACL)
 25. Murty S, Verga P, Vilnis L, McCallum A (2017) “Finer grained entity typing with typenet,” in Proc. 6th Workshop Automated Knowl. Base Construct, pp. 1–7
 26. Yao L, Riedel S, McCallum A (2013) Universal schema for entity type prediction. In Automatic Knowledge Base Construction Workshop at the Conference on Information and Knowledge Management
 27. Yaghoobzadeh Y, Schütze H (2016) Corpus-level fine-grained entity typing using contextual information. Proceedings of the Conference on Empirical Methods in Natural Language Processing
 28. Obeidat R, Fern XZ, Shahbazi H, Tadepalli P (2019) Description-based zero-shot fine-grained entity typing. In Proceedings of the 2019 Conference of the north American chapter of the Association for Computational Linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, volume 1 (long and short papers), pages 807–814. Association for Computational Linguistics
 29. Zhang T, Xia C, Lu C-T, Philip SY U (2020b) MZET: memory augmented zero-shot fine-grained named entity typing. In Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (online), December 8–13, 2020, pages 77–87. International Committee on Computational Linguistics
 30. Ren Y, Lin J, Zhou J (2020) Neural zero-shot fine-grained entity typing. In companion of the 2020 web conference 2020, Taipei, Taiwan, April 20–24, 2020, pages 846–847. ACM / IW3C2
 31. Ali MA, Sun Y, Li B, Wang W (2020) Fine-grained named entity typing over distantly supervised data based on refined representations. In The Thirty-F ourth AAAI Conference on Artificial Intelligence,

- AAAI 2020, The thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, pages 7391–7398. AAAI Press
32. Wu J, Zhang R, Mao Y et al (2022) Dealing with hierarchical types and label noise in fine-grained entity typing[J]. *IEEE/ACM Trans Audio, Speech, Lang Process* 30:1305–1318
 33. Dai H, Donghong D, Li X, Song Y (2019) Improving fine-grained entity typing with entity linking. In *Proceedings of the 2019 Conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP/IJCNLP 2019, Hong Kong, China, November 3–7, 2019*, pages 6209–6214. Assoc Comput Linguist
 34. Sun C, Li W, Xiao J, et al. (2021) Fine-grained chemical entity typing with multimodal knowledge representation[C]//2021 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 1984–1991
 35. Lee K, He L, Lewis M, Zettlemoyer L (2017) End-to-end neural coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*
 36. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. 2018. Deep contextualized word representations. In *proceedings of the 2018 conference of the north American chapter of the Association for Computational Linguistics: human language technologies (NAACL HLT 2018)*
 37. Radford A, Kim J W, Hallacy C, et al. (2021) Learning transferable visual models from natural language supervision[C]//international conference on machine learning. PMLR: 8748–8763
 38. Ren X, He W, Meng Q, Voss CR, Ji H, Han J (2016b) Label noise reduction in entity typing by heterogeneous partial-label embedding. In *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA, august 13–17, 2016*, pages 1825–1834
 39. Zhang S, Duh K, Van Durme B (2018) Fine-grained entity typing through increased discourse context and adaptive classification thresholds. In *proceedings of the seventh joint conference on lexical and computational semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 173–179
 40. Beltagy I, Lo K, Cohan A (2019) SciBERT: a pretrained language model for scientific text. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pp 3615–3620
 41. Li LH, Yatskar M, Yin D, Hsieh C-J, Chang K-W (2019) Visualbert: A simple and performant baseline for vision and language. *ArXiv preprint abs/1908.03557* (2019). <https://arxiv.org/abs/1908.03557>
 42. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, may 7–9, 2015, Conference Track Proceedings*
 43. Lin JC-W et al (2021) ASRNN: a recurrent neural network with an attention model for sequence labeling. *Knowl-Based Syst* 212:106548
 44. Lin JC-W et al (2020) Enhanced sequence labeling based on latent variable conditional random fields. *Neurocomputing* 403:431–440
 45. Shao Y et al (2021) Self-attention-based conditional random fields latent variables model for sequence labeling. *Pattern Recogn Lett* 145:157–164
 46. Sharma DK et al (2022) Explainable artificial intelligence for cybersecurity. *Comput Electr Eng* 103:108356
 47. Lin JC-W et al (2019) A bi-LSTM mention hypergraph model with encoding schema for mention extraction. *Eng Appl Artif Intell* 85:175–181
 48. Lv J et al (2023) Semi-supervised node classification via fine-grained graph auxiliary augmentation learning. *Pattern Recogn*:109301
 49. Azadifar S et al (2022) Graph-based relevancy-redundancy gene selection method for cancer diagnosis. *Comput Biol Med* 147:105766
 50. Nasiri E, Berahmand K, Li Y (2023) Robust graph regularization nonnegative matrix factorization for link prediction in attributed networks. *Multimed Tools Appl* 82(3):3745–3768

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.