Computer Communications 210 (2023) 1–9

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

A multimodal dual-fusion entity extraction model for large and complex devices

Weiming Tong [a], Xu Chu [b,*], Wenqi Jiang [b], Zhongwei Li [b]

[a] Laboratory for Space Environment and Physical Sciences, Harbin Institute of Technology, Harbin, China
[b] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, China

ARTICLE INFO

MSC:
68T01
68T50
68T99

Keywords:
Multi-source heterogeneity
Knowledge graphs
Multimodal
Dual-fusion
Entity extraction

ABSTRACT

In the context of large and complex devices with a multi-source heterogeneous data environment, the extraction of network device configuration related entity information from diverse modalities of the Internet of Things data is a crucial and fundamental step towards establishing a domain knowledge graph for global Zero Touch Provisioning of network devices. In this paper, we present a novel multimodal dual-fusion entity extraction model that serves as a foundation for an intelligent and efficient network device configuration process. Firstly, the multimodal data is encoded, followed by the ViLBERT pre-training model to obtain more feature information of entities for multimodal front-end fusion. Next, the attention weights of each modal feature are learned through a multilayer neural network classifier and probabilistic graphical model, facilitating multimodal back-end fusion and reducing information redundancy. Finally, entity recognition is accomplished by employing a cohesive memory module that extracts the essential parameters for device configuration. The simulation results demonstrate that the proposed model performs exceptionally well on the MSCOCO2017 public dataset and the SFZK-Dev data network device dataset, with $F1$ values of the comprehensive evaluation index of model quality at 94.65% and 96.94%, respectively, indicating high stability. Additionally, the link prediction metrics HITS@1 achieved accuracy levels of 58.21% and 68.04%.

1. Introduction

The Space Environment Simulation and Research Infrastructure (SESRI) represents a pivotal tool for simulating and researching the various environmental factors and their impacts on space exploration [1,2]. SESRI is a highly intricate system comprised of numerous interconnected systems. Within this system, a wide array of experimental and network devices exist, each belonging to different subsystems. These devices are distributed throughout SESRI, forming an Internet of Things (IoT) framework. The network encompasses vast volumes of structured, semi-structured, and unstructured data, including numerical, image, video, and text data. Consequently, this creates a heterogeneous data environment from multiple sources [3,4]. In order to effectively manage such systems with diverse and disparate data characteristics, developing comprehensive Knowledge Graphs (KGs) to configure network devices globally becomes crucial. KGs are semantic networks that reveal relationships between entities, and their mesh topology features enable efficient representation and visualization of topological relationships among network devices [5,6]. More precise and personalized configuration suggestions and recommendations can be generated by leveraging the semantic associations and data analysis capabilities of KGs. Zero Touch Provisioning (ZTP) can utilize the information extracted from KGs entities to automate configuration tasks, thus minimizing the need for manual intervention and physical contact.

In recent years, multimodality has gained increasing attention in KGs, Artificial Intelligence IoT (AIoT), and machine learning [7–10]. By integrating multimodal data using specific architectures and methodologies, we can effectively capture and utilize shared information across different modalities, resulting in improved machine comprehension and description of the device. This approach accurately extracts essential device configuration parameters, including location, network settings, and sensor thresholds. In the configuration process, leveraging multimodal data can offer a more comprehensive set of configuration information and contextual cues, thereby facilitating a deeper understanding of user requirements by the system. By integrating multimodal data fusion into the ZTP, an enhanced and more intelligent device configuration process can be achieved, leading to improved efficiency and effectiveness. Multimodal fusion architecture is classified into Joint, Coordinated, and Encode-Decode architectures [11]. Joint architecture projects unimodal representation into a shared semantic subspace, enabling multimodal feature fusion. The coordinated architecture comprises cross-modal similarity models and typical correlation analysis,

* Corresponding author.
E-mail address: 838700598@qq.com (X. Chu).

https://doi.org/10.1016/j.comcom.2023.07.026
Received 13 April 2023; Received in revised form 22 June 2023; Accepted 20 July 2023
Available online 28 July 2023

seeking correlations between modalities in the coordination subspace while maintaining the unique characteristics of each modality. Encode-Decode architecture is an intermediate representation that maps one modality to another, with each encoder and decoder encoding only one modality. Multimodal fusion methods are classified as either model-independent or model-based [12]. Model-independent fusion methods include feature fusion, which stitches together feature vectors from different modalities, and decision fusion, which integrates outputs of separately trained classifiers from different modal data. Model-based methods use deep learning models explicitly to solve multimodal fusion problems, including kernel-based methods, probabilistic graphical model methods, and neural network methods.

Knowledge extraction from multimodal data involves extracting high-quality factual representations to extract useful knowledge units through automated techniques, which primarily include three key elements: entities, relationships, and attributes. These elements lay the foundation for constructing the upper layer [13,14]. One fundamental task in knowledge extraction is entity extraction, also known as Named Entity Recognition (NER). The extraction of multimodal information mainly targets image and text modalities, and each data modality requires a different extraction approach. For images, features are typically extracted using stacked convolution and pooling operations. The output of a layer from an existing image classification or target detection model is used to represent the imaging modality, such as the output of the last CNN from the classification models ResNet and GoogLeNet, or the representation of a region from the target detection model R-CNN. For text, the focus is on the representation of basic language units, and a neural network is used to learn the language model to extract text features. Finally, some output vector from the neural network is utilized as the text representation.

In configuring large and intricate devices, crucial information, such as device characteristics, configuration descriptions, and compatibility details., is frequently stored in text and images, constituting a multimodal data source. The accuracy and comprehensiveness of extracting entities associated with network device configuration from this multimodal data source directly impact the efficiency of automated configuration. Current multimodal entity extraction methods treat each modality equally important and overlook the problem of an unbalanced distribution of instances across modal data. Diverse modal data provides more comprehensive information about a specific entity, with visual data enabling rapid determination of the device type and status and textual data facilitating quick identification of the device's IP address and performance, thus aiding in selecting appropriate configuration rules and parameters. Moreover, existing multimodal fusion techniques rely on a simple model structure and single-method hierarchical attention mechanism. The different modal information lacks an explicit interaction mode, which results in an incomplete exploitation of the complementary relationship between modal information.

This paper proposes a multimodal dual-fusion entity extraction model in a ZTP in the context of the IoT composed of large and complex devices to settle these issues. The proposed model leverages an embedding model to extract semantic information from each modal data and transform it into real-valued vectors. Pre-trained models and dual-fusion strategies are then applied to generate the complete semantic information of the data to extract information about network configuration-related entities from multimodal data, laying the foundation for building a more robust and intelligent configuration system and providing more comprehensive, accurate, and personalized configuration services. The main contributions are as follows:

- We have developed a comprehensive multimodal entity extraction model framework comprising three functional modules: an encoding module, a fusion module, and a joint memory module. The encoding module leverages Faster R-CNN to extract image feature information and a word dictionary model to extract text information. The fusion module combines a ViLBERT pre-trained model with a multilayer neural network classifier to learn the joint representation of static images and descriptive text. The joint memory module employs BiLSTM and CRF to capture entities' context and background knowledge, further enhancing the accuracy of entity recognition. Our experimental results demonstrate that the developed multimodal entity extraction model exhibits outstanding performance and can accurately extract device configuration-related entities from multimodal data.
- We propose a novel multimodal dual-fusion strategy for improving the KGs link prediction performance. Our approach comprises two parts: front-end fusion and back-end fusion. The front-end fusion leverages a dual-stream architecture to process image and text-coding vectors separately. We optimize the network depth for each modality independently and establish cross-modal connectivity at various depths to achieve effective information exchange between modalities. The back-end fusion involves calculating the weights of different modal features using a multilayer neural network classifier. Additionally, we model the conditional probability distribution of features based on a probabilistic graphical model to learn the attention weights of each modal feature and achieve complementarity of multimodal data. Experimental results demonstrate that our proposed multimodal dual-fusion strategy yields low-dimensional feature vectors that contain complete semantic information of multimodal data, thus improving the configuration template link prediction accuracy.

The remainder of this paper is organized as follows. Section 2 discuss related work,Section 3 presents the design of the multimodal entity extraction model and the dual fusion strategy. We report and discuss the experimental results in Section 4, followed by the conclusion in Section 5.

2. Related work

The proliferation of multimodal data in the IoT has laid the groundwork for leveraging multimodal models to enhance interactions and extract richer semantic information. Simultaneously, the combination of KGs technology to learn and extract critical information from multimodal data provides the possibility to realize AIoT global automated configuration. One such model is ViLBERT, proposed by Lu et al. [15], which enables joint representation learning of images and natural language by leveraging joint attention Transformer layers. However, the ViLBERT model is limited in its task complexity and the semantic information it extracts. To address these issues, Tan et al. [16] developed the LXMERT framework, which utilizes five pre-training tasks to enhance the connection between language and vision. Nevertheless, the LXMERT framework is computationally inefficient. Wang et al. [17] proposed the SimVLM model to achieve integrated utilization of multimodal data by constructing a sequence-to-sequence framework for end-to-end training of weakly aligned image-text pairs. Similarly, Xing et al. [18] introduced the KM-BART model, which acquires visual common sense knowledge through pre-training tasks to improve the model's performance. Zhao et al. [19] addressed the issue of missing partially modal data due to data sparsity by proposing a partially multimodal sparse coding model with adaptive similar structure regularization. Kang et al. [20] developed a consistent representation learning method for cross-modal retrieval that ensures the consistency of multimodal features using a local set of sparse regular terms. In the field of medical IoT, Gao et al. [21] proposed the CCJGSR model, which expresses test data through the sparse linear combination of training data and constraints observations from different modalities to share their sparse statements. However, the CCJGSR model does not fully leverage the complementary information among different features. Lastly, Sun et al. [22] proposed the VideoBERT model, which derives visual and language representations from the vector quantization of video data and speech recognition output results, respectively, to learn

higher-order semantic features through bi-directional joint distribution density.

Multimodal fusion is an approach that integrates multiple modalities of information to produce a consistent, unified output that enhances the robustness of predictions even when some modalities are absent. Several existing methods for multimodal fusion have different strengths and limitations. For instance, Han et al. [23] proposed a multi-view classification framework that jointly utilizes multiple views to enhance the reliability of classification. However, this method lacks interaction among the underlying multimodal data. Adir Solomon et al. [24] proposed a GRU predictive fusion model that embeds multimodal data into a common latent space and uses an attribute-specific attention mechanism to combine the embedded vectors into a fusion module. Nevertheless, the semantic integrity of each unimodal is not easily detectable at an early stage. Wang et al. [25] developed a complementary clothing- matching method that models visual relationships between items using neural graph networks and achieves multimodal information fusion through compatibility constraints. However, there is less interaction between the modal information. Wei et al. [26] proposed a multimodal cross-attention network for image and sentence matching that unifies association features within and between modalities for graphical matching computation. However, this method is less interpretable. Jin et al. [27] developed a recurrent neural network with an attention mechanism that fuses textual and social contextual features using an LSTM network and then combines them with image features. For homogeneous multimodal fusion, a method based on channel swapping was designed [28] to remove channels with less information in one modality and replace them with information from other modalities by sharing the same feature extraction network. However, this method is accompanied by data redundancy. Additionally, Zeng et al. [29] proposed an end-to-end deep fusion convolutional neural network to input 2D and 3D data into the network for feature extraction and fusion and then obtain a highly focused feature representation to effectively identify and localize target objects of different shapes.

Multimodal learning studies are guided by two essential principles, namely, the complementarity criterion and the consistency criterion. The primary objective is to develop models capable of processing and correlating information from multiple modalities. However, existing work in this area is characterized by a simplistic multimodal model structure, a single fusion method, and a lack of explicit interaction between different modalities. The current association between the modalities is only complementary, which only improves the robustness without fully exploiting the complementary relationship between modal information and fails to increase the amount of information. Motivated by recent studies, we propose a multimodal dual-fusion model that employs a pre-trained model for front-end decision fusion and utilizes different components such as a classifier, probabilistic graphical, and attention layers for back-end modal information fusion. Our model exhibits excellent performance in typical tasks.

3. Multimodal dual-fusion entity extraction model

This paper presents a multimodal dual-fusion entity extraction model aimed at accurately extracting crucial parameters of device configurations from the IoT environment, which is characterized by large and complex devices and encompasses a variety of heterogeneous data sources. The multimodal data utilized in our study consist of image and text modalities, each representing a different data source. Our model comprises three modules: encoding, fusion, and joint memory. The encoding module abstracts the semantic information embedded in the image and text modal data into real-valued vectors. The front-end fusion in the fusion module leverages the ViLBERT pre-training model to enable information interaction between the image and text modal data, resulting in mutual features between the two modal vectors. The back-end fusion calculates the sample weights of different modal feature vectors before fusion and class weights after front-end fusion by training a **Multilayer Neural Network Classifier (MNNC)** separately for different modal data. The model also learns the attention weights of each modal feature through probabilistic graph modeling to obtain multimodal low-dimensional feature vectors. Finally, the joint memory module enables accurate entity extraction. The extracted entities are then compared against pre-defined configuration templates, enabling the provision of technical assistance for the realization of automated configuration. Our multimodal dual-fusion entity extraction model architecture is presented in Fig. 1.

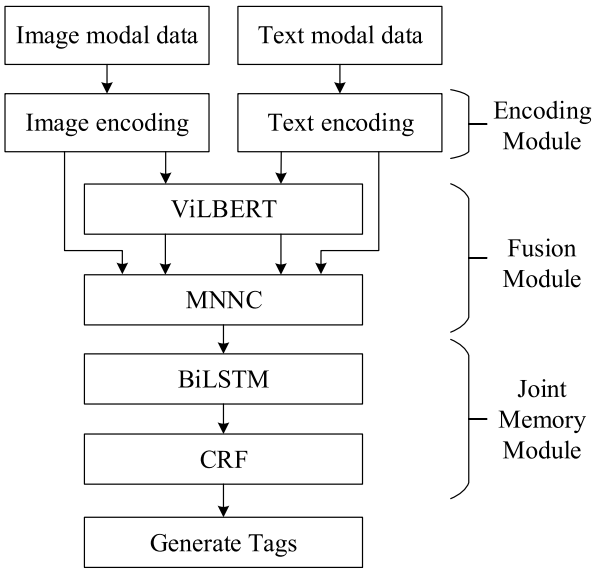


Fig. 1. Architecture of multimodal dual-fusion entity extraction model. The model consists of three main modules, including the encoding module, the fusion module, and the joint memory module.

3.1. Encoding module

To enable machines to process modal data effectively, it is necessary to represent the data as vectors. In the case of image data, a practical approach is to focus on feature regions of the image. The Faster R-CNN network model is an efficient solution that rapidly detects the feature regions. This model comprises three main components: the Extractor, Region Propose Network (RPN), and RoIhead. When an image is fed into the Faster R-CNN model, it undergoes an initial resizing operation to achieve a standardized size of 600 × 800 pixels. Subsequently, the resized image is forwarded through the backbone feature extraction network, an integral component of the model's architecture. The Extractor utilizes Convolutional Neural Networks (CNNs) to extract image features. Specifically, the VGG16 structure, which includes 13 convolutional layers and three fully connected layers, is employed to increase the network's depth and thereby enhance its performance. The RPN uses the extracted features to identify the regions of interest (ROIs) in the image, while RoIHead is responsible for fine-tuning and classifying the ROIs to determine their feature content and correct their position and coordinates. Finally, the feature vectors $p_1, p_2 \cdots p_m$ of the selected regions are obtained using average pooling and are represented by the <IMG> logo to denote the sequence start of image regions. These feature vectors serve as typical representations of the image for subsequent processing. Fig. 2 illustrates the overall framework diagram of the Faster R-CNN network model. The overall process of the Faster R-CNN algorithm can be divided into four steps.

Step 1: An ImageNet pre-trained model (VGG16) is used to initialize the RPN network and fine-tune the network parameters.

Step 2: Using the trained RPNs from Step 1 to generate proposals, a separate detection network is trained by Fast R-CNN, which the ImageNet pre-trained model also initializes.

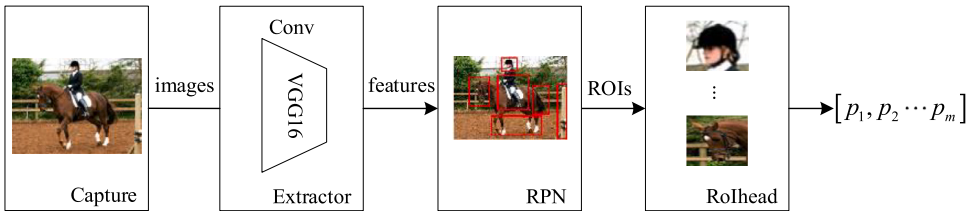


Fig. 2. Faster R-CNN network model framework. The model consists of four parts: Capture, Extractor, RPN, and RoIhead.

Step 3: Initialize the RPN training with the Fast R-CNN detection network, fix the share convolutional layers, and fine-tune only the RPN unique layers to achieve shared convolutional layers for both networks.

Step 4: Use the trained RPNs from Step 3 to generate proposals, keeping the shared convolutional layer fixed and fine-tuning the fully connected layer of the Fast R-CNN. The two networks share the same convolutional layers to form a unified network.

The text modality data is processed using the Stanford CoreNLP tokenizer to obtain a tokenized text sequence. The beginning of the text sequence is marked with the <CLS> tag, and <SEP> tags are inserted between different sentences to separate them. In order to effectively capture both semantic and syntactic information within the text, we leveraged the GloVe model to vectorize and encode the text data. The GloVe model initially constructs a Co-occurrence Matrix to quantify the frequency of word co-occurrences within the given text data. This matrix is then utilized to calculate the probability of co-occurrence between words. Subsequently, a loss function is formulated to minimize the squared difference between the discrepancy of two-word vectors and their corresponding co-occurrence probabilities and trained through iterative optimization using the defined loss function. Ultimately, a word vector representation is obtained, encapsulating the compositional structure of the text data and encoded as a vector, as denoted by $t_1, t_2 \cdots t_n$. Given that understanding text is generally more challenging for machines than understanding images, and that image inputs are typically high-dimensional feature vectors extracted by neural networks, the encoding depth required for text and image modalities is different. Therefore, in the early stages of our multimodal entity extraction model, we did not directly fuse the text modality data with the image modality data. Instead, we encoded the text modality data using a Transformer encoder for deeper encoding [30]. The Transformer encoder structure is illustrated in Fig. 3.

The Transformer block comprises multi-headed attention, a feed-forward neural network, residual connectivity, and layer normalization. Within this block, the multi-headed attention mechanism applies distinct linear transformations to the input features and calculates attention weights to derive multiple feature representations. The feed-forward neural network enables non-linear transformations on the output of the attention mechanism. Adding the input features to the features computed within the block incorporates residual connectivity. Additionally, layer normalization is employed to normalize the features, mitigating the issue of gradient disappearance and enhancing the training stability of the model. The Transformer encoder includes position encoding before processing the text vector. This process involves adding the position encoding to the input text vector T, which provides the relative position of each word within the sentence. The word representations are then obtained by adjusting the weight coefficient matrix based on the degree of association between words within the same sentence, as demonstrated in Eq. (1). Where q_T, k_T, v_T are the query vector, key vector, and value vector of each word in the input text, which are obtained by multiplying the input word vector with the respective weight matrix of the training; d_k is the dimension of the key vector.

$$E_T = softmax\left(\frac{q_T k_T^{(n)}}{\sqrt{d_k}}\right) v_T \tag{1}$$

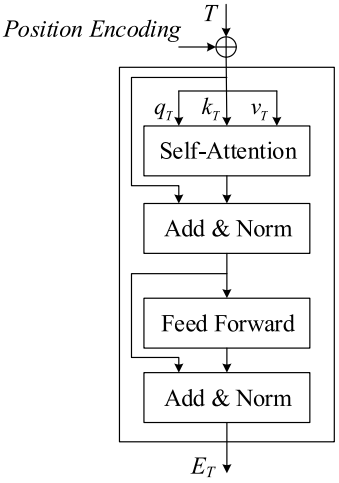


Fig. 3. Transformer encoder structure diagram. It consists of two sub-layers. The first one is a self-attention, and the second one is a fully connected feed-forward network.

3.2. Fusion module

The purpose of the fusion module is to merge text-encoded vectors and image-encoded vectors into a compact and unified vector, utilizing a property-specific attention mechanism. The fusion module comprises two key components: front-end fusion and back-end fusion.

Front-end fusion is accomplished through the utilization of the ViLBERT pre-training model. The ViLBERT pre-training model incorporates two primary pre-training tasks. The first is masked multimodal modeling, also called the reconstruction task. In this task, a portion of the image and text information is masked, and the model is then tasked with predicting the corresponding image region and text based on the masked input. In this paper, we design the masked multimodal modeling task by masking 15% of the input and using the remaining part to predict it, masking the image with 0 filling in 90% of the cases, and the region remains unchanged in 10%. The model predicts the distribution of image regions based on semantics and uses the minimization KL distance method to measure the relationship between the predicted distribution and the true distribution.

The second task is multimodal alignment prediction, also known as the matching task. Given the text and image inputs, the model must determine whether they align. In this paper, we design the multimodal alignment prediction task by using the output of the starting IMG token of the image feature sequence and the starting CLS token of the text sequence as the feature vectors for encoding the image and text streams, and then, the two feature vectors are synthesized into one vector by dot product (element-wise product) as the final overall representation. Finally, a linear layer predicts whether the image and text match.

The model leverages a dual-stream architecture to process both image and text encoding vectors, which allows for independent optimization of network depth for each modality and enables cross-modal

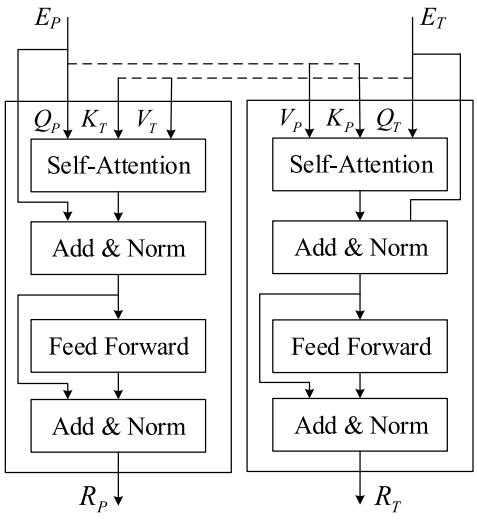


Fig. 4. Co-Attention Transformer layer structure diagram. It consists of a picture Transformer encoder and a text Transformer encoder, and each encoder exchanges key and value matrices with each other, thus forming Co-Attention.

connections at varying depths. The ViLBERT pre-training model is comprised of a Co-Attention Transformer layer (CoTRM) and a Transformer block (TRM) [31]. The architecture of the Co-Attention Transformer layer is illustrated in Fig. 4.

Similar to the Transformer encoder, the inputs to the Co-Attention Transformer layer are the image feature encoding vector E_P and the text feature encoding vector E_T. Initially, the query, key, and value matrices are derived from the weight matrix, which has been trained by the image and text processing streams, as outlined in Eq. (2).

$$\begin{cases} Q_P = E_P^{(n)} \times W_P^Q \\ K_P = E_P^{(n)} \times W_P^K \\ V_P = E_P^{(n)} \times W_P^V \\ Q_T = E_T^{(n)} \times W_T^Q \\ K_T = E_T^{(n)} \times W_T^K \\ V_T = E_T^{(n)} \times W_T^V \end{cases} \tag{2}$$

Subsequently, the key and value matrices of the image and text-processing streams are interchanged. The front-end fusion output of the image processing stream is obtained by leveraging its query matrix, and the key and value matrices from the text processing stream are utilized for computation. This process is represented in the image processing stream as the text attention under image conditions. Similarly, the front-end fusion output of the text processing stream is obtained by using its query matrix and the key and value matrices from the image processing stream for calculation, represented in the text processing stream as the image attention in the text condition. This approach allows for mapping information from one data modality to another, thus facilitating interaction between the image and text data. This, in turn, results in an output from both the image-processing stream and the text-processing stream that has mutually informative characteristics, thus enabling multimodal front-end fusion.

The back-end fusion procedure relies on classifiers separately trained for image and text modal data. A cross-entropy loss function L is introduced to address the mismatch between these modalities, as described in Eq. (3). The algorithmic steps for back-end fusion are as follows:

$$L = -\sum_{i=1}^{n} z_i \log\left(P_i\right) \tag{3}$$

Where n is the number of feature vectors, z_i is the correctness of the predicted category concerning the sample marker category, and P_i is the probability that the sample belongs to a certain category.

Step 1: The image feature encoding vector, which captures high-level semantic information through a deep neural network, and the text encoding vector, which incorporates contextual location information, are subjected to normalization before inputting into their respective classifiers. When the multimodal data is unfused, the image processing stream loss function L_P and the text processing stream loss function L_T are calculated. Following front-end fusion, the image vectors with text features are inputted to the image classifier, and the multimodal fused image processing stream loss function L_P^* is then calculated. Similarly, the text vectors with image features are inputted to the text classifier, and the multimodal fused text processing stream loss function is L_T^* calculated. The correlation coefficients between the intermodal loss functions are obtained by evaluating the values of each loss function, as shown in Eq. (4).

$$\rho_{LL^*} = \frac{\operatorname{cov}\left(L, L^*\right)}{\sigma_L \sigma_{L^*}} \tag{4}$$

Step 2: By integrating the front-end fusion features with the features from each modality and using a probabilistic graphical model, the multimodal dual-fusion entity extraction model is able to learn the conditional probability distribution of the front-end fusion features and the features from each modality, as shown in Eq. (5).

$$\begin{cases} P\left(R_P|E_T\right) = \|r-t\|_{\mathrm{A}} = (r-t)^{\mathrm{T}}\,\mathbf{A}\,(r-t) \\ P\left(R_T|E_P\right) = \|r-p\|_{\mathrm{A}} = (r-p)^{\mathrm{T}}\,\mathbf{A}\,(r-p) \end{cases} \tag{5}$$

Where R is the front-end fusion feature; E is the unfused feature, and $\mathbf{A}$ is the parameterization matrix.

Step 3: In the multimodal dependency relationship, implicit joint features affect different modalities. Attention weights for each feature vector of the image and text modality data are calculated based on this, expressed in Eq. (6).

$$\begin{cases} \alpha_{pi} = \arg\max_{\mathrm{A}} \operatorname{cov}\left(P\left(R_P|E_T\right), P\left(R_T|E_P\right)\right) - \lambda\rho_{LL^*} \\ \alpha_{ti} = 1 - \alpha_{pi} \end{cases} \tag{6}$$

Where $\operatorname{cov}(\cdot)$ is the dependence between multimodalities, ρ_{LL^*} is the correlation coefficient of the loss function, and $\lambda > 0$ controls the balance between loss functions and dependencies.

Step 4: To make fused decisions $x_i = \alpha_{pi} r_{pi} + \alpha_{ti} r_{ti}$ and obtain a new multimodal fused feature vector with reweighted importance of different modality features.

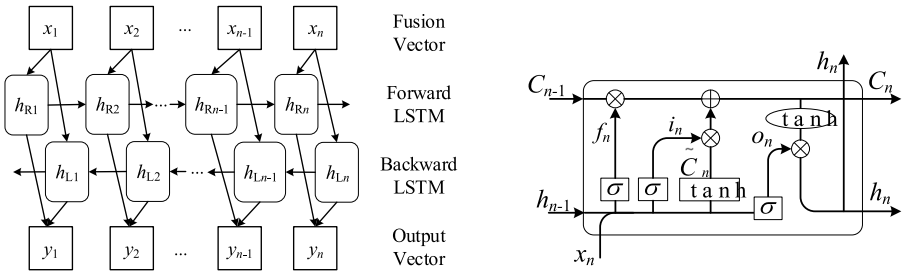
3.3. Joint memory module

The joint memory module comprises Bidirectional Long-Short Term Memory (BiLSTM) and Conditional Random Field (CRF). The purpose of this module is to capture the contextual information surrounding each instantaneous feature and thereby enhance the semantic understanding of the text for facilitating entity extraction [32].

The fundamental concept behind the BiLSTM is to feed a vector sequence into a sequence processing model, which comprises two LSTM networks operating in forward and backward directions. Subsequently, the outputs generated at each specific moment are merged. This enables each moment's output to be combined with the contextual information of its corresponding sequence, thereby augmenting the volume of information accessible to the network. The BiLSTM architecture is illustrated in Fig. 5. Eq. (7) represents the mathematical formulation for the LSTM structure.

$$\begin{cases} f_n = \sigma\left(W_{fx}x_n + W_{fh}h_{n-1} + b_f\right) \\ i_n = \sigma\left(W_{ix}x_n + W_{ih}h_{n-1} + b_i\right) \\ \tilde{C}_n = \tanh\left(W_{Cx}x_n + W_{Ch}h_{n-1} + b_C\right) \\ C_n = f_n * C_{n-1} + i_n * \tilde{C}_n \\ o_n = \sigma\left(W_{ox}x_n + W_{oh}h_{n-1} + b_o\right) \\ h_n = o_n * \tanh\left(C_n\right) \end{cases} \tag{7}$$

Where f_n, i_n, and o_n are the output results of the forgetting gate, input gate, and output gate, σ is the Sigmoid function, tanh is the



(a) BiLSTM structure diagram (b) Forward LSTM structure diagram

Fig. 5. BiLSTM architecture diagram.

hyperbolic tangent function, W is the corresponding weight matrix, b is the corresponding bias vector, and h_n is the output result of the LSTM.

As illustrated in Fig. 5, the output h_{Ri} of the ith forward LSTM is computed based on the first i terms, whereas the output h_{Li} of the ith backward LSTM is determined based on the n-i to nth terms. The BiLSTM output is obtained by concatenating the outputs of the forward and backward LSTMs, which are computed as depicted in Eq. (8).

$$y_n = \left[h_{Ln}, h_{Rn}\right] \tag{8}$$

Multimodal entity extraction involves extracting entities from multiple data sources, such as text and images. While BiLSTM is capable of obtaining information from the distal end of the vector sequence, it falls short in handling the correlation between adjacent elements. On the other hand, CRF can effectively leverage the correlation between adjacent elements to obtain an optimal prediction sequence. Therefore, adding a CRF layer to the output of BiLSTM can compensate for its limitations. To calculate the score of a certain labeled sequence B, we sum the weights of all the feature functions that meet the given condition. The CRF output sequence can be expressed using Eq. (9).

$$B = \arg\max_{B \in B_Y} \left(\sum_{i=1}^{n} \omega \cdot g_i \left(b_{i-1}, b_i, y_i\right) \right) \tag{9}$$

Where B_Y is all possible labeling sequences, g_i is the state feature function, and ω is the weight of the feature function.

The final multimodal dual-fusion entity extraction process is shown in Fig. 6.

4. Simulation and results analysis

4.1. Dataset

In this study, we utilized the publicly available MSCOCO2017 dataset as the benchmark to evaluate our proposed model. This dataset was collected using Amazon Mechanical Turk and contains annotation information for images, including category and location data, as well as semantic text descriptions of images. The dataset consists of 118,287 pairs for the training set and 5,000 for the validation set. Additionally, this study creates the SFZK-Dev dataset using network devices in the SESRI. The dataset includes images and configuration descriptions of network devices. The dataset was randomly partitioned into three sets, with 8,000 pairs allocated for training, 1,000 for validation, and 1,000 for testing purposes. We utilize the BIOES (Begin, Inside, Outside, End, Single) annotation scheme for textual data and adopt the bounding box approach for image data.

4.2. Evaluation metrics, experimental environment and parameter configuration

Given that entities represent the fundamental components of a KGs, the thoroughness, precision, and recall of their extraction have a direct bearing on the overall quality of the knowledge base. Therefore, entity extraction is considered to be a pivotal and foundational step in the process of knowledge extraction. This study employs various metrics to evaluate the quality of the model, including recall rate (R), precision rate (P), and the comprehensive $F1$ score. Additionally, link prediction metrics such as HITS@n are also employed to evaluate the performance of the model in the configuration template matching task. The calculation methodologies for each of these evaluation metrics are presented in Eqs. (10) and (11).

$$\begin{cases} P = \frac{e_r}{e} \times 100\% \\ R = \frac{e_r}{E} \times 100\% \\ F1 = \frac{2PR}{P+R} \times 100\% \end{cases} \tag{10}$$

Where e_r is the number of correctly identified entities, E is the total number of entities, and e is the number of identified entities.

$$\text{HITS@}n = \frac{1}{|S|} \sum_{i=1}^{|S|} \Pi \left(rank_i \le n\right) \tag{11}$$

Where S is the set of triples, $|S|$ is the number of sets of triples, $rank_i$ is the link prediction ranking of the ith triple, and Π is the indicator function (if the condition is true then the function value is 1, otherwise it is 0).

The experimental setup is based on the Linux operating system, with an Intel i7-12700H CPU, an NVIDIA GeForce RTX 3060 GPU, and 32 GB of memory. The Tensorflow deep learning environment of Python 3.7+CUDA 11.0.3 is utilized. The Adam optimizer is employed to train the model with a learning rate of 0.001. LSTM_dim is set to 200 and the Dropout method is utilized to prevent overfitting. The batch size is set to 64, and the training process is conducted over 20 epochs.

4.3. Simulation results

In this paper, we evaluate the performance of four entity extraction models respectively, Unimodal-Picture (U-P), Unimodal-Text (U-T), Multimodal-Single Fusion (M-SF), and Multimodal- Dual Fusion (M-DF). Among them, single fusion in M-SF is implemented using ViLBERT pre-training model.

In the context of model generalizability, this paper employs the publicly available MSCOCO2017 dataset as the foundation. A comparative analysis of precision and recall rates across different models is conducted, as shown in Fig. 7. The results indicate that the recall rate of the multimodal model can reach nearly 100%, while the unimodal text flow model achieves a recall rate of 90%. In contrast, the recall rate of the unimodal picture flow model is only around 85%. It can be inferred that certain entities exhibit more prominent features within images, which can provide richer semantic information for entity extraction. Furthermore, the precision rate of entity extraction using a multimodal model is consistently higher across the recall range, exhibiting an improvement of approximately 20% compared to unimodal entity extraction. Thus, the use of multimodal models for entity extraction is demonstrated to be beneficial. To further evaluate the performance of the model, a comprehensive evaluation metric of model quality

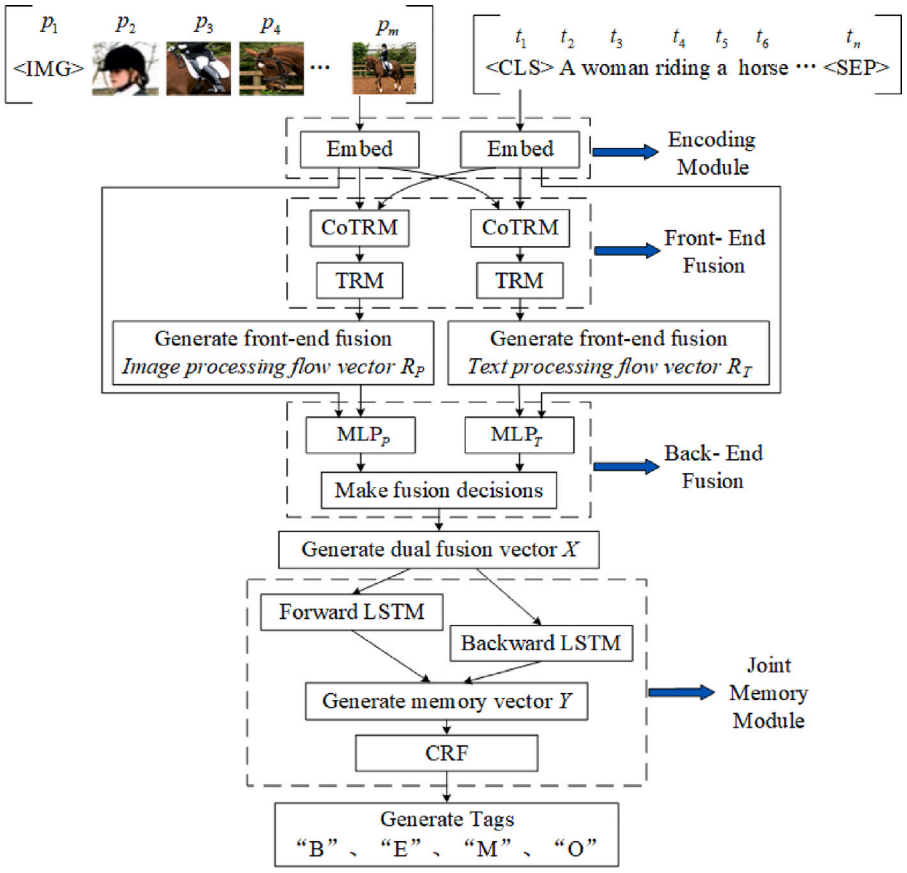


Fig. 6. Flow chart of multimodal dual fusion entity extraction. It is first through the encoding module, followed by front-end fusion, then back-end fusion, and finally, the tags are output through the joint memory module.

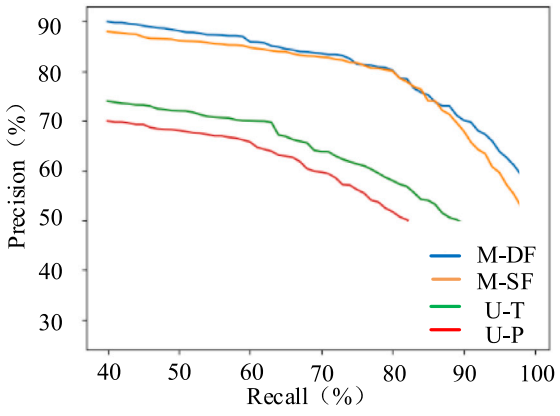


Fig. 7. Plot of precision and recall for the four models.

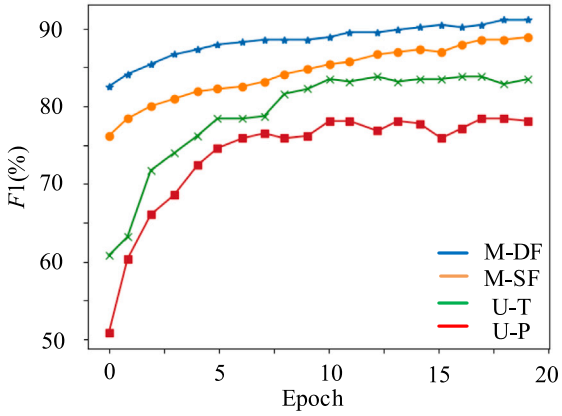


Fig. 8. Plot of $F1$ value update with the number of Epoch.

is introduced, which incorporates both precision and recall rates to calculate the $F1$ value. Fig. 8 illustrates the number of model epochs with $F1$ value updates. The multimodal model exhibits a higher starting point of $F1$ value, around 80% when the number of epochs is small. As the number of epochs increases, the $F1$ value increases gradually and stabilizes at around 90%. However, the multimodal dual fusion model consistently outperforms the other models, with a lead of 5% at the beginning of the epoch. In contrast, the unimodal model exhibits a lower starting point of $F1$ value at the beginning of the epoch, but after several epochs and iterative updates, the $F1$ value increases rapidly and stabilizes at a similar level to the multimodal model, yet remains below it. Based on these findings, it can be concluded that the multimodal model is able to extract more entity descriptions than the unimodal model. Moreover, the multimodal dual fusion model can quickly learn from diverse data and achieve good performance with high stability at the early stages of the epoch. This paper assesses the performance of each entity extraction model by utilizing the HITS@n (where $n = 1, 3, 10$), and the results are presented in Table 1. The findings demonstrate that the multimodal model outperforms the unimodal model across all stages. Notably, the multimodal model with dual-fusion exhibits a significant advantage over the single fusion model, achieving an accuracy increase of approximately 8% at HITS@1. Therefore, the dual-fusion model delivers superior accuracy when the link prediction

Table 1
Link prediction evaluation metrics HITS@n of four models (n = 1, 3, 10)

Mode	Hits@1	Hits@3	Hits@10
Unimodal-Picture(U-P)	27.43%	68.93%	76.78%
Unimodal-Text(U-T)	36.55%	71.82%	80.80%
Multimodal-Single Fusion(M-SF)	50.60%	72.42%	83.62%
Multimodal- Dual Fusion(M-DF)	**58.21**%	**75.18**%	**85.07**%

Table 2
Optimal performance of multimodal dual fusion model for each evaluation metric with two datasets.

Evaluation Metrics	MSCOCO2017	SFZK-Dev
Precision	94.38%	97.21%
Recall	94.92%	96.68%
$F1$	94.65%	96.94%
HITS@1	58.21%	68.04%

ranking is small, enabling rapid and precise information retrieval and enhancing the computational efficiency of the model.

The optimal performance of the proposed multimodal dual-fusion entity extraction model for each evaluation metric under the two datasets is shown in Table 2. The results show that the proposed model performs well on the datasets. Notably, the model achieved the best performance on the SFZK-Dev dataset, which relies on artificial prior knowledge to process crucial information, leading to higher scores across all evaluation metrics.

5. Conclusion

KGs play a crucial role in storing, analyzing, and reasoning about information of network configuration in the IoT, comprising a vast array of complex devices. They are indispensable for achieving ZTP. Leveraging multimodal data allows for a more comprehensive understanding of configuration information and enables intelligent decision-making during the configuration process. This paper proposes a novel multimodal dual-fusion entity extraction model that integrates KGs and multimodal data for ZTP. This model facilitates explicit interaction among different information modalities and harnesses the complementary relationship between them, thereby serving as a foundation for realizing an automated and highly efficient configuration process. The following conclusions are drawn: (1) ViLBERT pre-training model is used to obtain more feature information for entities. Two independent encoding modules process textual and visual information, achieving independent optimization of each modality's network depth and cross-modal connections at different depths. (2) A multimodal double fusion strategy is proposed. Front-end fusion achieves information interaction between image and textual entities through an attention mechanism, while back-end fusion utilizes a multi-layer neural network classifier to calculate attention vectors for each modality and stage. Then, based on the probability graph model, attention vector weight conditional probability distribution is calculated, and multiple attention vectors are fused according to the weights, thereby more effectively fusing multimodal information. (3) A joint memory module based on BiLSTM and CRF is constructed to obtain entity context and background knowledge, improving entity recognition accuracy. (4) The proposed multimodal double fusion entity extraction model is trained and tested on the public dataset MSCOCO2017 and the established SFZK-Dev dataset. Simulation results show that the model performs excellently in various evaluation metrics. Device configuration key information can be accurately obtained on the SFZK-Dev dataset and matched with configuration templates.

Future research will focus on enhancing computational efficiency while maintaining model accuracy and exploring methods to create lightweight models that can be rapidly deployed in ZTP for efficient and automatic global deployment of IoT devices.

Declaration of competing interest

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.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by the "Space Environment Simulation and Research Infrastructure" of National Major Science and Technology Infrastructure Construction Project (Big Science Project) of National Development and Reform Commission (Fagai High Technology paper No. 24 in 2017), and was supported by the Preliminary study on "Space Environment Simulation and Research Infrastructure" of National Major Science and Technology Infrastructure of Funding Projects for Fundamental Scientific Research Operation Fees of Central Universities (30620180135).

References

[1] H.P. Jiang, J.L. Liu, H.F. Hao, Sesri 300mev proton and heavy ion accelerator, J. Phys. Conf. Ser. (2019) 012081.
[2] Y. Jiuwei, Y. Yaqing, L. Minbang, C. Wenjun, Vibration of scanning magnet for space environment simulation and research infrastructure, High Power Laser Part. Beams 05 (2021) 91–95.
[3] X. Zhang, D. Xue, J. Zheng, The transmission and parse technology of multi-source heterogeneous data based on opc ua in intelligent manufacturing, Mech. Electr. Eng. Technol. 01 (2021) 1–7.
[4] F. Zhang, J. Yang, C. Sun, X. Guo, Research on multi-source heterogeneous data fusion technology for complex information system, China Meas. Test 07 (2020) 1–7+23.
[5] F. Zhao, H. Sun, L. Jin, Structure-augmented knowledge graph embed ding for sparse data with rule learning, Comput. Commun. 159 (2020) 271–278.
[6] X. You, Y. Ma, Z. Liu, Representation method of cooperative social network features based on node2vec model, Comput. Commun. 173 (2021) 21–26.
[7] Q. Zheng, H. Wen, M. Wang, Visual entity linking via multi-modal learning, Data Intell. 1 (2022) 1–19.
[8] Y. Liu, H. Li, A. Garcia-Duran, Mmkg: multi-modal knowledge graphs, in: European Semantic Web Conference, 2019, pp. 459–474.
[9] M. Wang, G. Qi, H. Wang, Richpedia: A comprehensive multi-modal knowledge graph, in: Joint International Semantic Technology Conference, 2019, pp. 130–145.
[10] X. Zhu, Z. Li, X. Wang, Multi-modal knowledge graph construction and application: A survey, IEEE Trans. Knowl. Data Eng. (2022).
[11] C. Zhang, Z. Yang, X. He, Multimodal intelligence: Representation learning, information fusion, and applications, IEEE J. Sel. Top. Sign. Proces. 3 (2020) 478–493.
[12] T. Baltrušaitis, C. Ahuja, L.P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 2 (2018) 423–443.
[13] R. Xu, T. Liu, L. Li, Document-level event extraction via heterogeneous graph-based interaction model with a tracker, 2021, arXiv preprint arXiv:2105.14924.
[14] R.L. Logan IV, S. Humeau, S. Singh, Multimodal attribute extraction, 2021, arXiv preprint arXiv:171111118.
[15] J. Lu, D. Batra, D. Parikh, ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. (2019).
[16] H. Tan, M. Bansal, LXMERT: Learning cross-modality encoder representations from transformers, 2019, arXiv preprint arXiv:1908.07490.
[17] Z. Wang, J. Yu, A.W. Yu, SimVLM: Simple visual language model pre-training with weak supervision, 2021, arXiv preprint arXiv:2108.10904.
[18] Y. Xing, Z. Shi, Z. Meng, KM-BART: Knowledge enhanced multimodal bart for visual commonsense generation, 2021, arXiv preprint arXiv:2101.00419.
[19] Z. Zhao, H. Lu, C. Deng, Partial multi-modal sparse coding via adaptive similarity structure regularization, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 152–156.
[20] C. Kang, S. Xiang, S. Liao, Learning consistent feature representation for cross-modal multimedia retrieval, IEEE Trans. Multimed. 3 (2015) 370–381.
[21] Z. Gao, Y. Yang, M.R. Khosravi, Class consistent and joint group sparse representation model for image classification in internet of medical things, Comput. Commun. 166 (2021) 57–65.

[22] C. Sun, A. Myers, C. Vondrick, VideoBERT: A joint model for video and language representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7464–7473.
[23] Z. Han, C. Zhang, H. Fu, Trusted multi-view classification, 2021, arXiv preprint arXiv:2102.02051.
[24] A. Solomon, B. Shapira, L. Rokach, Predicting application usage based on latent contextual information, Comput. Commun. 192 (2022) 197–209.
[25] R. Wang, J. Wang, Z. Su, Learning compatibility knowledge for outfit recommendation with complementary clothing matching, Comput. Commun. 181 (2022) 320–328.
[26] X. Wei, T. Zhang, Y. Li, Multi-modality cross attention network for image and sentence matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10941–10950.
[27] Z. Jin, J. Cao, H. Guo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 795–816.
[28] Y. Wang, W. Huang, F. Sun, Deep multimodal fusion by channel exchanging, Adv. Neural Inf. Process. Syst. (2020) 4835–4845.
[29] H. Zeng, J. Luo, Construction of multi-modal perception model of communicative robot in non-structural cyber physical system environment based on optimized bt-svm model, Comput. Commun. 181 (2022) 182–191.
[30] A. Vaswani, N. Shazeer, N. Parmar, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).
[31] H. Xu, K. Zhang, Y. Tian, Image caption algorithm based on ViLBERT and BiLSTM, Comput. Syst. Appl. 11 (2021) 195–202.
[32] T. Xie, J. Yang, H. Liu, Chinese entity recognition based on bert-BiLSTM- CRF model, Comput. Syst. Appl. 07 (2020) 48–55.