Intelligent and Converged Networks ISSN 2708-6240
2022, 3(2): 161−172 DOI: 10.23919/ICN.2022.0010

From computer vision to short text understanding: Applying similar approaches into different disciplines

Jiayin Lin*, Geng Sun, Jun Shen, David E. Pritchard, Ping Yu, Tingru Cui, Dongming Xu, Li Li, and Ghassan Beydoun

Abstract: With the development of IoT and 5G technologies, more and more online resources are presented in trendy multimodal data forms over the Internet. Hence, effectively processing multimodal information is significant to the development of various online applications, including e-learning and digital health, to just name a few. However, most AI-driven systems or models can only handle limited forms of information. In this study, we investigate the correlation between natural language processing (NLP) and pattern recognition, trying to apply the mainstream approaches and models used in the computer vision (CV) to the task of NLP. Based on two different Twitter datasets, we propose a convolutional neural network based model to interpret the content of short text with different goals and application backgrounds. The experiments have demonstrated that our proposed model shows fairly competitive performance compared to the mainstream recurrent neural network based NLP models such as bidirectional long short-term memory (Bi-LSTM) and bidirectional gate recurrent unit (Bi-GRU). Moreover, the experimental results also demonstrate that the proposed model can precisely locate the key information in the given text.

Key words: neural network; natural language processing; deep learning

1 Introduction

With the development of Internet and communication technologies such as IoT and 5G, the forms of online information resources develop from single form (e.g., pure text or pure picture) to diversification (e.g., in a multimodal form). For example, a single online learning course can contain textual information (e.g., the discussion from students), audio-video information (e.g., the recorded lecture content), and images (e.g., the attached slides)[1]. Alternatively, in the area of e-commerce, a list of goods on Amazon or Taobao can have text, images, and even short video for its description[2]. Hence, solely processing single format of the information is no longer adequate enough to fully understand the online resources. Such changes pose new challenges for the intelligent multimodal information processing tool towards various big data application areas, including e-learning, digital health, traffic information systems, etc.

However, the researches on information processing and fusion for handling different formats of the

• Jiayin Lin is with the College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350108, China. E-mail: jy.lin@fjnu.edu.cn.
• Geng Sun, Jun Shen, and Ping Yu are with the School of Computing and Information Technology, University of Wollongong, Wollongong 2500, Australia. E-mail: {gsun, jshen, ping}@uow.edu.au.
• David E. Pritchard is with the Research Lab of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: dpritch@mit.edu.
• Tingru Cui is with the School of Computing and Information Systems, University of Melbourne, Melbourne 3010, Australia. E-mail: tingru.cui@unimelb.edu.au.
• Dongming Xu is with the UQ Business School, University of Queensland, Brisbane 4000, Australia. E-mail: d.xu@business.uq.edu.au.
• Li Li is with the Faculty of Computer and Information Science, Southwest University, Chongqing 400715, China. E-mail: lily@swu.edu.cn.
• Ghassan Beydoun is with the School of Information System and Modelling, University of Technology Sydney, Sydney 2007, Australia. E-mail: ghassan.beydoun@uts.edu.au.
∗ To whom correspondence should be addressed.
Manuscript received: 2021-12-05; revised: 2022-03-16; accepted: 2022-05-06

information are still relatively isolated even though many efforts had been made independently or diversely. For example, there is a lack of a robust universal model, which can handle the processing task of different types of information (i.e., multimodal information). Hence, multiple models are generally involved in a processing system to handle and interpret multimodal information. In the prior work of managing open educational resource in medicine discipline[3], three different models were used to handle text, video stream, and pictures. Ultimately, a universal and reuseable model or a single model only requiring little adjustment for different types of information, can greatly reduce the complexity of the overall system.

When tracing back the final objective of tasks of natural language processing (NLP) or computer vision (CV) problems, we can see that most of these studies aim to construct a model to understand the content of a piece of text or a picture. For the convenience of training process, the content of a picture or text is always represented in the form of digitalized numbers (either integer or decimal). The training process of the model is conventionally based on a certain mathematic concept to manipulate these numbers and narrow down the gap between the prediction and the pre-defined ground truth. In short, the CV problem and NLP problem share a similar goal and the data are represented similarly in both CV and NLP tasks. The highlights of the paper are to discuss the common underlying mechanisms between NLP and CV models, and investigate the possibility of using CV techniques to solve NLP problems. In this study, we propose to investigate the following question: " Can similar machine learning approaches, models, ideas, or techniques be used in different disciplines, such as CV and NLP, where certain application areas require holistic handling on multimodal information forms?"

In order to answer the above question, we initiate that the effort of using the mainstream solutions in CV area could be made to understand the content of the short texts. Two different tweet datasets with different training objectives are used in this study. The contributions of this paper are summarized below:

(1) The demonstration of how to model an NLP problem from the perspective of computer vision based on the real-world datasets.

(2) Proposal of a pilot convolutional neural network solution to understand the content of the short texts.

(3) Comprehensive experiments with detailed analysis to demonstrate the merits of our proposed solution and the possibility of using generic model to process multimodal information.

The remainder of this paper is organized as follows: Related works of the mainstream solutions in CV and NLP will be firstly introduced and discussed in Section 2. In Section 3, the proposed solution to understand the content of the short text will be presented. The experiment part will be demonstrated in Section 4, which includes the details of evaluation metrics, datasets, experimental settings, and description of the baselines and analysis of the experimental results. This paper will be concluded in Section 5, together with the plan of future research.

2 Related work

2.1 CNN in computer vision

With the outstanding performance in extracting different granularities of spatial information, the convolutional neural network (CNN) based models have become the mainstream solutions to various computer vision tasks[4]. To name a few, for the task of image classification, the AlexNet[5], which consists of five convolutional layers, can distinguish one thousand different objectives. For the task of object detection, R-CNN[6], Fast R-CNN[7], and Grid R-CNN[8] are well-adopted region based CNN with the idea of the bounding box to locate target objects. Moreover, besides recognizing objects, the CNN-based models can also work as a tracker in the task of object tracking. One representative object tracking model is FCNT[9], which takes advantage of the feature map from the model VGG[10]. Multi-hierarchical independent correlation filter is also used for visual tracking[11]. As the network goes deep, different types of features (such as edges, shape, and so on) are extracted and modelled successively, but the network will face the challenge of vanishing gradients. Hence, the ResNet[12], which consists of 152 layers, is designed to maintain the

robustness of the deep model. All these works demonstrate the outstanding non-linear transforming ability of CNN mechanism, which can be used to interpret the content of the complex data or signals in format of pixels. Hence, we are interested in verifying the research question that, as the representation of a sequence of text is similar to a picture, whether the mainstream CV models work as good as the mainstream NLP models when dealing with an NLP problem?

2.2 RNN and CNN in NLP

In NLP, constrained by "rules" (such as the grammar and idioms), a meaningful sentence is commonly composed with specific patterns. When interpreting the meaning of a piece of text, a sequence of a bunch of words is usually more important than the individual words themselves. Due to the significance of modelling temporal/time-series information, the recurrent neural network (RNN) based model can be regarded as dominant for various NLP tasks in the recent years. For example, the combination of RNN and conditional random field (CRF) based framework has been regarded as the optimal choice for the tasks of sequence labelling and information extraction in various application areas[13]. Similarly, for the task of text classification and text generation, most of the existing solutions are based on the famous RNN framework and its variants[14]. Hence, it is prudent to raise a research question that, with proper modifications, can CNN model capture and interpret temporal patterns as RNN model does and consequently be applied in the NLP task?

2.3 CNN in NLP and RNN in CV

From the viewpoint of mathematics, the goals of both the CNN and RNN networks are to project the given input to the required output through complex non-linear transformation. In recent years, many prior studies have been successfully using the CNN model to solve NLP problems or RNN model to solve CV problems. In Ref [13], CNN is used to generate word embeddings by extracting character-level semantic information such as prefix and suffix. TextCNN[15] is designed for the task of sentence classification, and this work demonstrated that a simple CNN with little hyperparameter tuning has the potential to achieve excellent results on multiple benchmarks. In the area of CV, the combination of CNN and RNN is used for image classification in Refs. [16, 17]. The LSTM-C framework is used in incorporating the knowledge from external sources in Ref. [18] to address the issue of predicting novel objects in image captioning task. However, these studies do not do and explore the relationship between these two areas into further depth and do not analyse whether it could be feasible to have a universal framework or model for both areas.

3 Model design

In this section, we describe the design of the proposed CNN-based framework for interpreting the tweet content. First, from a high-level perspective, we introduce the overall architecture of this framework. Next, how the research problem is formulated in this study will be demonstrated.

3.1 Architecture framework

The proposed framework contains two components: the upstream component and the downstream component. The general architecture of the proposed framework is shown in Fig. 1. The upstream component is the pre-trained language model; it is used to transform the raw text input into the dense embeddings. The downstream component is the task-specific model, which takes in the dense embeddings and produces the final predictions for a specific NLP task. In this study, we only focus on applying CV ideas to the downstream component, whereas designing the upstream component is beyond the scope of this paper as many researchers have been carried out in this field. However, to get comprehensive experiment results, different pre-trained language models will be compared and discussed in this paper later.

3.2 Problem formulation

To clearly formulate the task of tweet content understanding, we would have the following definitions.

3.2.1 Upstream component

In this study, the pre-trained language model L is used to generate dense word embeddings $\boldsymbol{e}$ with dimension m. For each word w in a tweet $\boldsymbol{t} = (w_1, w_2, \ldots, w_i)$, the

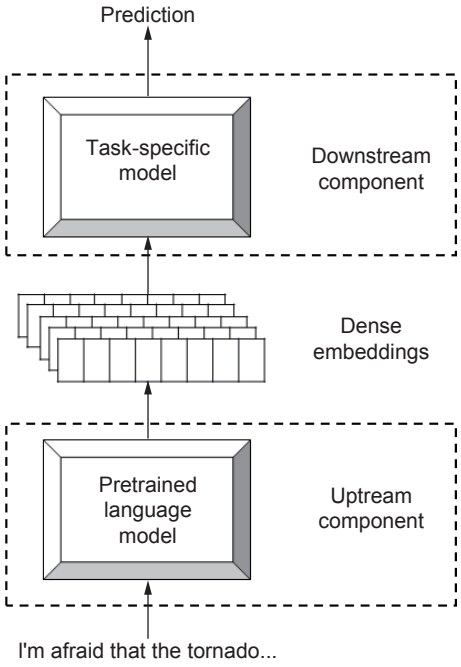

Fig. 1 Framework architecture.

mapping procedure can be formulated as $\boldsymbol{L}(w_i) => \boldsymbol{e}_i$, $\boldsymbol{e}_i \in R^m$, where $\boldsymbol{e}_i$ is an m-dimentional real value vector, R^m is the set of these vectors.

3.2.2 Tweet representation

In this study, a tweet is represented in the form of 2D "figure" $\boldsymbol{P}$ by stacking all embeddings together, such procedure is formulated as $\boldsymbol{P} = Stack(e_1, e_2, \ldots, e_i)$. To prevent any information loss in this procedure, the order of the embeddings is kept the same with the original text sequence. The illustration of this process is shown in Fig. 2.

3.2.3 Downstream component

Given different NLP tasks have different goals, the designs of the downstream component can vary from task to task. Usually the recurrent neural network and its variants are used to capture temporal information. However, a CNN-based downstream model is used in this study. This downstream component takes in generated text "figure" $\boldsymbol{P}$ and makes the final prediction $\boldsymbol{y}$. Two tweet datasets are used in this study, one for disaster prediction and another for sentiment analysis. For a disaster prediction task, the goal is to analyze the content of a given tweet $\boldsymbol{t}$ and predict whether a tweet is about real disaster or not. For this task, defining the ground-truth as $y_{\text{disaster}} \in \{0, 1\}$, where 1 is for disaster and 0 is for not about the disaster. For the sentiment analysis task, the goal is to find out whether a given tweet $\boldsymbol{t}$ contains a certain sentiment or not, and what its sentiment is. For this task, defining the ground-truth as $y_{\text{sentiment}} \in \{0, 1, 2\}$, where 0 is for neutral (no obvious sentiment involved in a tweet), 1 is for negative sentiment, and 2 is for positive sentiment. Hence, together with the above two definitions, the goal of the downstream component is to learn the following function $\mathcal{F}$:

$$\mathcal{F}(\boldsymbol{P}) \Rightarrow y \tag{1}$$

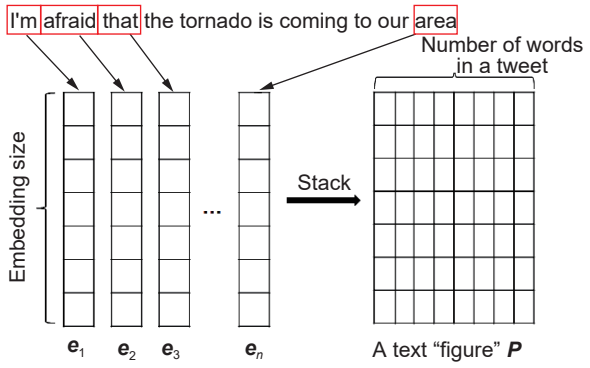

Fig. 2 Organization of a piece of text "figure".

Inspired by Ref. [15] and the idea of the n-gram model[19], we have carefully designed a CNN-based network to interpret the text content. The generated 2D text figure is scanned by multiple kernels k_1−k_i serval times to extract semantic information. Different kernels have different widths d_i but share the same height h. The kernel height h is equal to the word embedding size. With these settings, kernels can summarize the information of number of d_i successive words at each step during the first convolutional operation, such procedure is similar to generating n-gram samples. The following convolutional operations extract different levels of granularity of information in the same way as it has been frequently used in the CV area. After a set of successive convolutional operations, pooling operation and fully connected layers are applied to summarize all extracted information and produced the final predictions. The illustration of the network structure of the proposed CNN-based downstream component is shown in Fig. 3.

3.2.4 1D bounding box

Moreover, in this study, despite making the proposed model understand tweet content, we also investigate

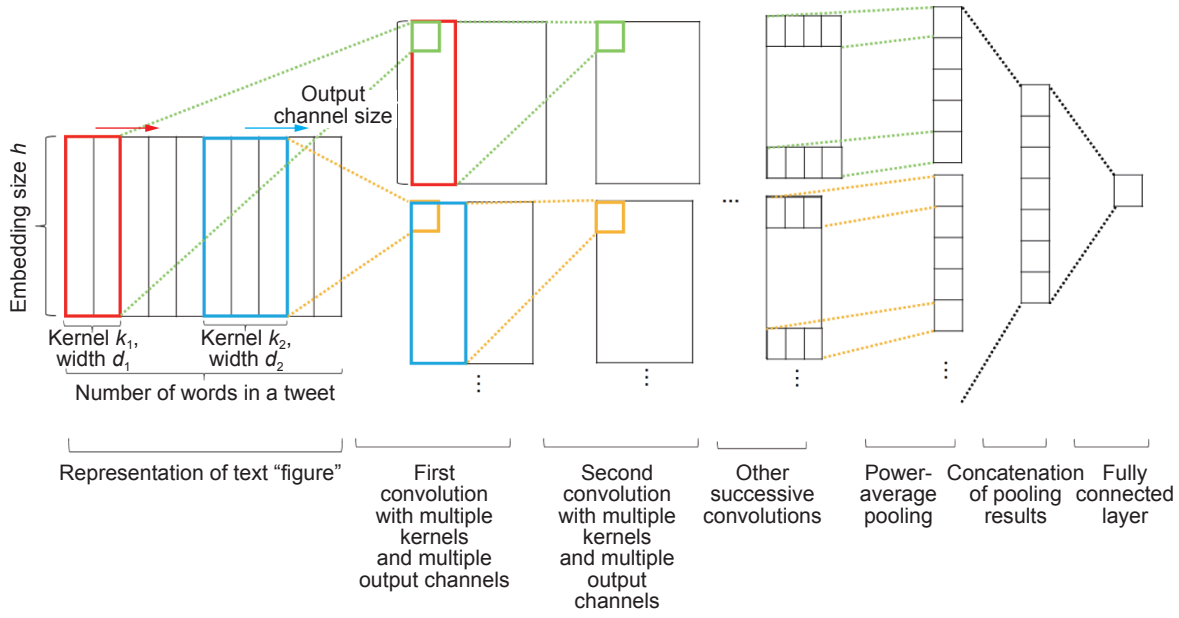

Fig. 3 Network structure of CNN-based downstream component.

whether the proposed model can identify which phrases or words in the tweet would express the key information. In another word, we want the model to be able to identify which words or phrases could express a negative or positive sentiment for a given tweet. Inspired by the utility of the bounding box in object detection, a 1D "bounding box"※ is used to mark the key information. Hence, the proposed model will not only classify a given text to one predefined category but also highlight which words/phrases express the key information that triggers the model make this decision. The main difference between the bounding box used in this study and the one used in a common CV task is that our bounding box only has one dimension (horizontal). The example of the bounding box in both areas is shown in Fig. 4. In Fig. 4, the left part is the bounding box of the object detection results of cats; the right part is the bounding box of the sentiment analysis result of negative feeling. The 1D bounding box is formulated as $\boldsymbol{B}(c_{\text{index}}, \text{l})$, the first element is the center index of the selected sequence of text and the second element is the length of selected text. The loss function to measure the prediction and the ground-truth of the bounding box is formulated as

※ The working manner and optimization process of the proposed bounding box is different from the one in computer vision. We merely use a similar idea to identify the wanted information.

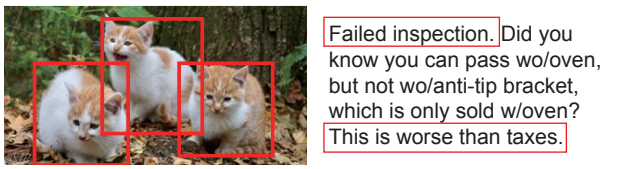

Fig. 4 Example represntation of bounding box in CV (left) and NLP (right).

$$Loss_B = MSE(c_{\text{index}} - \widehat{c_{\text{index}}}) + MSE\left(1 - \widehat{1}\right) \quad (2)$$

where the *MSE* is the mean square error loss. Difference between the predicted centre and the ground-truth centre is measured by the first term, and the difference between the predicted length of key words/phrase and the ground-truth length is measured by the second term.

4 Experiment and analysis

The experimental details are demonstrated and analysed in this section, including the introduction of the datasets, the used evaluation metrics, the baselines, the experimental setting, and the analysis of the results.

4.1 Dataset

There are two short text datasets used in the experiments; and both are collected from the Twitter platform and open to the public‡. Both datasets are

‡ The dataset about disaster prediction was created by the company figure-eight and originally shared on https: //www.figure-eight.com/data-for-everyone/. The second dataset about sentiment analysis was extracted from https: //appen.com/resources/datasets/.

short informal text collected from the Twitter platform. The reasons of choosing these two datasets are: (1) Handling informal text to prove that our proposed model is general enough to handle complex NLP problem. (2) Using short text to demonstrate the outstanding information extracting ability of our proposed model.

The first dataset contains more than 10 thousand tweets, some of which talking about the real disaster events. The ratio of disaster-related against non-disaster-related is 43:57. This dataset contains the raw text of each tweet, keyword from the tweets, and location information where a tweet was sent from. The keyword and location information may be blank; we only utilise raw text information to train the model. The second dataset contains 30 thousand tweets with or without sentiment information. The ratio of the number of neutral against positive against negative sample is 41:31:28. This dataset contains the raw text of each tweet and the text fragment that support the tweet's sentiment. Similarly, only the raw text is used to train the model. The raw text is collected directly from the Twitter platform and has not been pre-processed yet. The example of these two datasets is shown in Table 1.

4.2 Evaluation metrics

In order to reflect the model performance from different perspectives, three different types of evaluation metrics are used in the experiment. The first evaluation metric is accuracy (*Acc*), formulated as Eq. (3), which directly reflects the proportion of the correct predictions produced by each model. However, for imbalanced distributed ground-truth, this metric might not be suitable for comparing the effectiveness of the models[20].

$$Acc = \frac{total\ number\ of\ correct\ prediction}{total\ number\ of\ prediction} \tag{3}$$

The second evaluation metric used in this study is the area under curve (AUC) value. AUC value reflects the ability of a model to distinguish different types of information (i.e., for task one about whether it is a disaster or not, for task two about whether they are different sentiments). When dealing with the multi-class classification task (task of sentiment analysis), the one-versus-one strategy is used in the experiment. AUC value is the area under the receiver operating characteristic (ROC).

The last evaluation metric is the F-score, formulated as Eq. (4), which is the harmonic mean of the recall score and the precision score. Because of the trade-off between recall and precision, we cannot conclude a good model merely based on high Recall score or high precision score. Hence, using F-score is a better choice for model comparison.

$$F_1 = 2 \times \frac{recall \times precision}{recall + precision} \tag{4}$$

4.3 Baseline

In the experiments, different pre-trained models are used to investigate the effectiveness of applying the CV solution to the NLP problem. Specifically, for the upstream component, we involve the following pre-trained language models:

(1) Word2Vec[21]: This model has high optimization efficiency, but can only model the local semantic information within the pre-defined window.

(2) GloVe[22]: This model combines the merits of LSA[23] and Word2vec. It uses the co-occurrence matrix to model the local and global semantic information at the same time.

(3) Bert[24]: Bert and its variants dynamically model the semantic information. As indicated in the original study that a multitask fine-tuning approach could be used to train the model, and this would boost the performance even further.

Table 1 Data sample of the two datasets.

Dataset	Information type	Information
Disaster prediction	Raw text	*I-77 Mile Marker 31 to 40 South Mooresville Iredell Vehicle Accident Congestion at 8/6 1:18 PM*
	Keyword	*accident*
	Location	*North Carolina*
Sentiment analysis	Raw text	*A little happy for the wine jeje ok it`sm my free time so who cares, jaja i love this day*
	Text fragment	*A little happy*

For the downstream component, we compare the following models' effectiveness in understanding tweet content:

(1) Bi-GRU: Bi-directional gated recurrent unit neural network.

(2) Bi-LSTM: Bi-directional long-short-term-memory neural network.

(3) The proposed CNN-based model in our paper.

4.4 Experimental setup

In this study, all the models are implemented using PyTorch framework[25]. The pretrained language models Bert and Word2Vec are implemented through Transformer§ and Gensim¶, respectively; the pre-trained GloVe model is reproduced through its pretrained word vectors☆. Six different kernels with 128 output channels are used in the proposed CNN-based model. ReLU is used as the activation function for each convolutional output, and there are four successive convolutional layers in total. The dimension number for the hidden layer of Bi-LSTM and Bi-GRU is set to 256. The sigmoid function is used to produce the final prediction for disaster prediction task, while the softmax function is used to produce the final prediction for sentiment analysis task. All the other settings are strictly stick to the original work, or we directly use the default settings of the PyTorch framework. The early-stop mechanism is applied to all the training processes to prevent overfitting.

Before using the language model to convert the tweet content to dense vectors, the raw text is pre-processed through several NLP data cleaning and normalizing stages, ranging from removing the stop-words and lemmatization to removing URLs and emojis.

4.5 Experiment results and discussions

Table 2 illustrates the effectiveness comparison of different downstream components for two different NLP tasks. Table 3 reports the effectiveness of different upstream components. As we have obtained similar results from two datasets, we only present the experimental results of the task disaster prediction in Table 3. The demonstration of the 1D bounding box to locate key information is shown in Table 4.

§https: //huggingface.co/transformers/index.html
¶https: //radimrehurek.com/gensim/#
☆https: //nlp.stanford.edu/projects/glove/

Table 2 Downstream component comparison on the tasks of sentiment analysis and disaster prediction.

Task	Model	*Acc*	F_1	AUC
Disaster prediction	Bert + proposed model	**0.8238**	**0.8169**	**0.8798**
	Bert + Bi-GRU	0.8108	0.8044	0.8698
	Bert + Bi-LSTM	0.8172	0.8128	0.8695
Sentiment analysis	Bert + proposed model	**0.8573**	**0.8354**	**0.8998**
	Bert + Bi-GRU	0.8465	0.8273	0.8800
	Bert + Bi-LSTM	0.8159	0.7927	0.8543

Table 3 Upstream component comaprion on the task of disaster prediction.

Model	*Acc*	F_1	AUC
Bert + proposed model	**0.8238**	**0.8169**	**0.8798**
GloVe + proposed model	0.8173	0.8145	0.8723
Word2Vec + proposed model	0.6852	0.6762	0.7271
Bert + Bi-GRU	**0.8108**	**0.8044**	**0.8698**
GloVe + Bi-GRU	0.7667	0.7598	0.8097
Word2Vec + Bi-GRU	0.6819	0.6760	0.7348
Bert + Bi-LSTM	**0.8172**	**0.8128**	**0.8695**
GloVe + Bi-LSTM	0.8107	0.8039	0.8588
Word2Vec + Bi-LSTM	0.6770	0.6625	0.7283

4.5.1 Effectiveness of the proposed CNN model

According to the results demonstrated in Table 2, we can easily conclude that, with the same Bert upstream component, the proposed CNN-based downstream component shows competitiveness comparing to mainstream NLP solutions in all criteria for two different tasks (highlighted in bold text). For the task of disaster prediction, Bi-LSTM model slightly outperforms the Bi-GRU model, while for the task of sentiment analysis, the Bi-GRU model greatly outperforms the Bi-LSTM model. We would argue such improvement is produced by the structure difference between the LSTM cell and GRU cell. According to the original work of LSTM[26] and GRU[27], the LSTM tends to remember longer semantic information than GRU. Hence, we consider that for different tasks, to involve too much information during the modelling procedure will not always improve the model performance.

The task of disaster prediction needs to understand the whole tweet to predict whether the given tweet is about disaster or not. It is hard to infer whether a tweet is about a disaster or not merely based on a short text segment or phrase. As shown in the following example:

Table 4 Demonstration of 1D bounding box on the task of sentiment analysis.

Raw text	Selected text	Ground-truth key word	Sentiment
Grrr..stupid internet connection ruined a great scrabble game	grrr..stupid	Grrr..stupid internet connection	negative
listening to the best days of your life by kellie pickler	listening to the best days	listening to the best days of your life	positive
awesome! All deserved I'm sure. Miss the Crabs games	awesome	awesome	positive
Are you going to hate being around my baby?	are you going to hate being	hate	negative
awww I love me some charlies we are enjoying some lucky food LOL	awww i love me some	love	positive
NICE! Got any that are indexed that you want to unload? I need a few.	nice! got any that	NICE!	positive
Still gutted that man utd lost	Still gutted that man utd lost	Still gutted that man utd lost	negative
i miss him ALOT but im not gonna talk to him, i HOPE	i miss him alot but im not gon na talk to him	miss	negative
Starting to spoil my pug since her brother Max passed away on Tuesday. We miss him.	starting to spoil my pug	We miss him.	negative
Just try to do your best. I hope you don't get laid off.	try to do your best.	Just try to do your best. I hope you don't get laid off.	positive

"All residents asked to ' shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected."

Remembering more words or longer text sequence is helpful to understand the context of the whole tweets. Hence, for such task, the LSTM-based model is better than GRU-based model.

As for the task of the sentiment analysis, in most cases, sentiment is just one part of a tweet content. As a negative sentiment tweet shown in the following example, a user posts a certain event and expresses how he/she feels about it: " Grrr..stupid internet connection ruined a great scrabble game."

Interpreting the sentiment of a user relies on short text sequence or phrase. Inferring from the longer text sequence might negatively affect the model performance, as the phrase " a great scrabble game" shows positive sentiment in the above example, which contains negative sentiment. Hence, with the same settings, the Bi-GRU model greatly outperforms the Bi-LSTM model in the second task.

The above long-short-sequence modelling problem can be avoided by using the proposed CNN-based model. When using the CNN-based model, we can flexibly control the length of word sequence to be modelled at each step through setting the kernel size at the first convolutional layer. The configuration of the kernel sizes can be determined on domain knowledge or pilot experiments.

4.5.2 Importance of upstream language model

We also investigate the framework performances in using different language models in the upstream component. From Table 3, we can conclude that the frameworks with pretrained Bert model outperform the ones with GloVe model or Word2Vec model (highlighted in bold text). As mentioned in Section 4.3, the Bert model can capture more semantic information. A better language model indicates the downstream component can access more useful information. From the CV perspective, a better language model can generate " higher resolution text figure", which is critical for mining details from the " text figure". As argued in Ref. [28], tweets generally contain more information per character, likely a result of Twitter-specific abbreviations and a less consistent writing style. Hence, to better interpret the tweet content, it is necessary to use powerful language model for maximizing the retention of semantic information.

4.5.3 Locating key information

Making use of the second dataset, we also further investigate whether our proposed model can identify and locate the key information (i.e., keywords). As mentioned in Section 3, a 1D-bounding box is designed to select the keywords (for an NLP task, such job is usually accomplished by using attention mechanism). Due to the space limitation, in this paper, we only randomly present five positive and five negative results as shown in Table 4. The column of "Ground-truth key

word", is showing the labelled ground-truth words that contain sentiment information. The column of "Selected text" is showing the key words selected by the bounding box. We can clearly see that our proposed model can not only understand the sentiment meaning of the tweet content but also identify which words express such meaning. The only drawback of our bounding box is that it is prone to selecting longer text than the ground truth (by comparing the second and the third column). We attribute it to the common situations that longer text could contain more information and would be more supportive of the bounding box to make a decision.

5 Conclusion and future work

In this study, we investigate the feasibility of designing a generic model to solve the multimodal information problem. A CNN-based model and a further bounding box are proposed to demonstrate that, with proper adjustments, the mainstream solutions from CV area can also be used to solve NLP problems with different goals. With proper configurations, the proposed CNN-based model can yield a better generalization ability than RNN-based model. Based on the experiment results, we discover the significance of the language model in modelling textual information. The experimental results have also shown that, for the task of short text understanding, our proposed solution has competitive performance comparing to mainstream NLP solutions such as Bi-LSTM and Bi-GRU. Moreover, the proposed model shows satisfactory performance in locating key information.

In the future, firstly, we will continue investigating the effectiveness of the proposed solution for long text understanding. To better design such a generic model for processing multimodal information, we will also investigate in applying CNN-based solution to solve other forms of information such as audio signals. In the meantime, we will also try to use NLP solutions reversely to solve other forms of information.

Acknowledgment

This work was supported by the Australian Research Council Discovery Project (No. DP180101051) and Natural Science Foundation of China (No. 61877051).

References

[1] J. Chen, H. Li, W. Wang, W. Ding, G. Y. Huang, and Z. Liu, A multimodal alerting system for online class quality assurance, in *Proc. 20th International Conference on Artificial Intelligence in Education*, Chicago, IL, USA, 2019, pp. 381–385.

[2] C. Lynch, K. Aryafar, and J. Attenberg, Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank, in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 541–548.

[3] B. Zhao, S. Xu, S. Lin, X. Luo, and L. Duan, A new visual navigation system for exploring biomedical open educational resource (OER) videos, *Journal of the American Medical Informatics Association*, vol. 23, no. e1, pp. e34–e41, 2016.

[4] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, Deep learning vs. traditional computer vision, in *Proc. 2019 Computer Vision Conference* (*CVC*), Las Vegas, NV, USA, 2019, pp. 128–144.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Region-based convolutional networks for accurate object detection and segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[7] R. Girshick, Fast R-CNN, in *Proc. IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1440–1448.

[8] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, Grid R-CNN, in *Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 7355–7364.

[9] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, Fully-convolutional siamese networks for object tracking, in *Proc. 2016 European Conference on Computer Vision*, Amsterdam, the Netherlands, 2016, pp. 850–865.

[10] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409.1556, 2014.

[11] S. Bai, Z. He, Y. Dong, and H. Bai, Multi-hierarchical independent correlation filters for visual tracking, in *Proc. 2020 IEEE International Conference on Multimedia and Expo* (*ICME*), London, UK, 2020, pp. 1–6.

[12] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVRP*), Las Vegas, NV, USA, 2016, pp. 770–778.

[13] Z. Zhai, D. Q. Nguyen, and K. Verspoor, Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition, arXiv preprint arXiv: 1808.08450, 2018.
[14] D. Pawade, A. Sakhapara, M. Jain, N. Jain, and K. Gada, Story scrambler-automatic text generation using word level RNN-LSTM, *International Journal of Information Technology and Computer Science* (*IJITCS*) , vol. 10, no. 6, pp. 44–53, 2018.
[15] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv: 1408.5882, 2014.
[16] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, CNN-RNN: A unified framework for multi-label image classification, in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2285–2294.
[17] Y. Guo, Y. Liu, E. M. Bakker, Y. Guo, and M. S. Lew, CNN-RNN: A large-scale hierarchical image classification framework, *Multimedia Tools and Applications*, vol. 77, no. 8, pp. 10251–10271, 2018.
[18] T. Yao, Y. Pan, Y. Li, and T. Mei, Incorporating copying mechanism in image captioning for learning novel objects, in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 5263–5271.
[19] B. Roark, M. Saraclar, and M. Collins, Discriminative n-gram language modeling, *Computer Speech & Language*, vol. 21, no. 2, pp. 373–392.
[20] F. J. Valverde-Albacete and C. Peláez-Moreno, 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox, *PloS ONE*, vol. 9, no. 1, p. e84217, 2014.
[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv: 1301.3781, 2013.
[22] J. Pennington, R. Socher, and C. D. Manning, Glove: Global vectors for word representation, in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, Qatar, 2014, pp. 1532–1543.
[23] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.
[24] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805, 2018.
[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., PyTorch: An imperative style, high-performance deep learning library, in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS 2019*), Vancouver, Canada, 2019, pp. 8026–8037.
[26] F. A. Gers, J. Schmidhuber, and F. Cummins, Learning to forget: Continual prediction with LSTM, in *Proc. 1999 Ninth International Conference on Artificial Neural Networks* (*ICANN 99*), Edinburgh, UK, 1999, pp. 850–855.
[27] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint arXiv: 1406.1078, 2014.
[28] G. Neubig and K. Duh, How much is said in a tweet? A multilingual, information-theoretic perspective, in *Proc. 2013 AAAI Spring Symposium: Analyzing Microtext*, Palo Alto, CA, USA, 2013, pp. 32–39.

Jun Shen received the PhD degree from Southeast University, China in 2001. He held positions at Swinburne University of Technology in Melbourne and University of South Australia in Adelaide before 2006. He is an associate professor in the School of Computing and Information Technology at University of Wollongong in Wollongong, NSW of Australia, where he had been the head of Postgraduate Studies and chair of School Research Committee since 2014. He is a senior member of three institutions: IEEE, ACM, and ACS. He has published more than 200 papers in journals and conferences in CS/IT areas. His expertise includes computational intelligence, bioinformatics, cloud computing, and learning technologies including MOOC. He has been an editor, PC chair, guest editor, and PC member for numerous journals and conferences published by IEEE, ACM, Elsevier, and Springer. He was also a member of ACM/AIS Task Force on Curriculum MSIS 2016. His publications appeared at *IEEE Transactions on Learning Technologies*, *Briefs in Bioinformatics*, *International Journal of Information Management*, and many others.

Geng Sun was a research fellow in the School of Computing and Information Technology, University of Wollongong, Australia. He received the PhD degree from University of Wollongong in 2018 with Examiners' Commendation for Outstanding Thesis. He held an Australian Postgraduate Award and also received an award of Outstanding Self-Financed Students Aboard from Chinese Government. Since he commenced pursuing the master degree in University of Wollongong, he has published more than 30 papers in leading journals, including *IEEE Transactions*, and conferences in e-learning area, with a focus on collaborative learning previously and micro learning now. He is also an ad hoc reviewer for several high-ranked e-learning journals and conferences such as *IEEE Transactions on Learning Technologies*, *Interactive Learning Environments*, and International Conference on Computer Supported Collaborative Learning.

Jiayin Lin received the BEng degree from University of Electronic Science of Technology of China in 2011, the MS degree in computer science from University of Melbourne in 2015, and the PhD degree from University of Wollongong in 2021. He used to work as an algorithm engineer at Alibaba Digital Media & Entertainment Group. He is currently a lecturer in the College of Computer and Cyber Security at Fujian Normal University, China. His research interests are machine learning, deep learning, and recommender system.

David E. Pritchard received the BS degree from California Institute of Technology and the PhD degree from Harvard University, working with Professor Daniel Kleppner, who soon thereafter came to MIT, bringing Pritchard with him as a postdoctoral researcher. He joined the faculty of the Department of Physics in 1970. He has many honors, including the Broida Prize of the American Physical Society. He is a member of the National Academy of Sciences and is a fellow of the American Academy of Arts and Sciences. He carried out pioneering experiments on the interaction of atoms with light that led to the creation of the field of atom optics. His demonstration of the diffraction of a beam of atoms by a grating made of light waves opened the way to studies of the diffraction, reflection, and focusing of matter waves, similar to those with light waves. He has applied atom optics to basic studies of quantum theory, to new methods for studying the properties of atoms, and to the creation of devices such as the atom interferometer and atom wave gyroscope. He is the creator of the largest engineering and science homework tutorial system in the US — MasteringPHysics.com, MasteringBiology.com, etc. He has also published more than 50 papers on MOOCs and studying learning in online courses (and online cheating, which is anti-learning).

Tingru Cui received the PhD degree from National University of Singapore in 2014. Her research interests are interdisciplinary in nature, exploring the human-AI interaction, business and learning analytics, social media, crowdsourcing, and digital innovation. She uses a range of methodologies, including data mining, natural language processing, lab experiment (eye tracking, EEG), case study, survey, and econometric analysis in her projects. She focuses on understanding the interplay between human, information, and technology, including theories and applications related to behaviour, business, and social impact of IT. She is an enthusiastic researcher with over 50 peer reviewed publications, including 30 A*/A publications.

Ping Yu is the director of Centre for Digital Transformation in the School of Computing and Information Technology, University of Wollongong, Australia. She also leads the Group of Digital Health and Digital Aged Care in Smart Infrastructure Facility at UOW. She is an elected fellow of Australasian College of Health Informatics and an associate editor of *BMC Medical Informatics and Decision Making*. She is one of the earliest researchers in the world to pursue mobile health research. She pioneered the digital health adoption research in residential aged care early in 2003. Sustained leadership has seen her team to contribute a substantial body of cutting-edge knowledge in digital aged care, i.e., electronic nursing documentation and digital solutions for continence care, wound care, and medication management. She also led the digital transformation of patient appointment system in a primary health care clinic in Australia. Her recent digital transformative effort is focused on secondary use of electronic health records, applying artificial intelligence and data analytics to optimise health and aged care services. Some exemplar projects are (1) application of theory-based, mobile and social media technology for digital transformation of consumer lifestyle and chronic disease management; (2) hospital patient journey mapping and treatment process optimisation; (3) ontology driven, multi-agent technology to improve aged care services; and (4) observational study and data analytics to optimise resource utilisation in health and aged care.

Dongming Xu is a senior lecturer in business information systems at the UQ Business School. Her research focuses on the confluence of information technology use and information technology innovation to reach a deep understanding of how information systems are used and how information systems influence the society.

Li Li is a professor in the Faculty of Computer and Information Science at Southwest University, China. She received the PhD degree from Swinburne University of Technology, Melbourne, Australia in the field of computing. She was the postdoctoral researcher at CSIRO in Canberra, Australia. She secured two Natural Science Foundation Project of China (NSFC) grants and two state-level grants (Natural Science Foundation Project of Chongqing). Her research interests include machine learning and artificial intelligence in education. She has authored/co-authored 2 books and over 120 papers in international journals and conferences such as KDD, ICDM, DASFAA, PAKDD, ICWS, and ICSOC. She is the member of IEEE.

Ghassan Beydoun is a professor of information systems at the School of Information System and Modelling, University of Technology Sydney. He is also currently the deputy head of School (Research). He received the PhD degree from the University of New South Wales, Australia. His principal area of research interest is modelling focusing on development methodologies for distributed intelligent systems and decision support systems. His research is funded by the Australian Research Council and a number of industry collaborators, focusing on a decision making applications including: disaster management, supply chain management, health policies, learning systems, and cloud computing. He also serves on the editorial board of *Information Systems Frontiers*, *Journal of Theoretical and Applied Electronic Commerce Research*, *International Journal of Intelligent Information Technologies*, and *Journal of Software*. His latest publications appeared in *IEEE Transactions of Software Engineering*, *Journal of Human Computer Studies*, *European Journal of Information Systems*, *Journal of Systems and Software*, and others.