Multi-modal Generative Adversarial Networks for Traffic Event Detection in Smart Cities

Qi Chen[a], Wei Wang[a,*], Kaizhu Huang[b], Suparna De[c] and Frans Coenen[d]

[a] *Department of Computer Science and Software Engineering, Xi'an Jiaotong Liverpool University, China*
[b] *Department of Electrical and Electronics Engineering, Xi'an Jiaotong Liverpool University, China*
[c] *Computer Science and Networks Department of Digital Technologies, University of Winchester, United Kingdom*
[d] *Department of Computer Science, University of Liverpool, United Kingdom*

ARTICLE INFO

Keywords:
Multi-modal learning
Semi-supervised learning
Traffic event detection
Generative adversarial network
Smart transportation
Deep learning

ABSTRACT

Advances in the Internet of Things have enabled the development of many smart city applications and expert systems that help citizens and authorities better understand the dynamics of the cities, and make better planning and utilisation of city resources. Smart cities are composed of complex systems that usually process and analyse big data from the Cyber, Physical, and Social worlds. Traffic event detection is an important and complex task in smart transportation modelling and management. We address this problem using semi-supervised deep learning with data of different modalities, e.g., physical sensor observations and social media data. Unlike most existing studies focusing on data of single modality, the proposed method makes use of data of multiple modalities that appear to complement and reinforce each other. Meanwhile, as the amount of labelled data in big data applications is usually extremely limited, we extend the multi-modal Generative Adversarial Network model to a semi-supervised architecture to characterise traffic events. We evaluate the model with a large, real-world dataset consisting of traffic sensor observations and social media data collected from the San Francisco Bay Area over a period of four months. The evaluation results clearly demonstrate the advantages of the proposed model in extracting and classifying traffic events.

1. Introduction

Today, technologies from the Internet of Things (IoT) have been widely used to address challenges that modern cities face, e.g., traffic congestion, air pollution, energy consumption and public safety. Intelligence Transportation Systems (ITS), as an instance of smart city applications, aim to discover knowledge from traffic related data collected from a city environment for efficient management of transportation and mobility in a city. For example, Lv et al. exploited historical traffic flow data for traffic prediction (Lv et al., 2015); Song et al. engaged GPS records with millions of anonymous users for human mobility prediction (Song et al., 2016); Anantharam et al. collected social media data for traffic event detection (Anantharam et al., 2015). However, most of these existing studies collect and analyse data from either the physical world (Lv et al., 2015; Song et al., 2016) or social world (Anantharam et al., 2015; Gu et al., 2016).

Smart city is a typical Cyber-Physical-Social (CPS) system, which usually collects, processes and analyses data of different types and modalities. It is common that different sources may publish incomplete data in different modalities about the same physical phenomenon. Obviously, data from different sources should complement and knowledge discovered should reinforce each other, e.g., a traffic anomaly that is not inferred from traffic sensor observations might be clearly explained by a number of tweets. Nevertheless, the challenging problem is how to design a unified framework for processing such multi-modal data which differs greatly in the level of granularity and semantic meaning. Our work aims to exploit data of different modalities while complementary to each other to extract trustworthy knowledge and improve classification performance.

With the rapid development of ITS, the generated traffic data collected from loop sensors, GPS, cameras and social media, is exploding. Researchers believe that we have entered the era of big data transportation (Lv et al., 2015). Much of the transportation research focus has shifted towards the processing of massive amounts of data continuously generated within a city environment. However, most of the existing studies only process data of single modality and require a large amount of labelled data, which is usually not practical in real-world, big data applications. This inspires us to design a semi-supervised learning framework for traffic event detection with the rich amount of unlabelled data and the extremely limited amount of labelled one.

Deep learning is a popular paradigm in the machine learning family. It allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction, and is able to discover intricate structures from natural data in its raw forms without the need for sophisticated feature engineering and tuning (LeCun et al., 2015). Studies based on deep models have significantly improved the state-of-the-art in various intelligent transportation related applications, such as traffic flow prediction (Lv et al., 2015), traffic event detection (Zhang et al., 2018), and smart parking (Valipour et al., 2016). Deep learning techniques are also considered as good candidates for knowledge fusion and integration, e.g., the re-

*Corresponding author

✉ qi.chen@xjtlu.edu.cn. (Q. Chen); wei.wang03@xjtlu.edu.cn. (W. Wang); kaizhu.huang@xjtlu.edu.cn. (K. Huang); Suparna.De@winchester.ac.uk. (S. De); Coenen@liverpool.ac.uk. (F. Coenen)

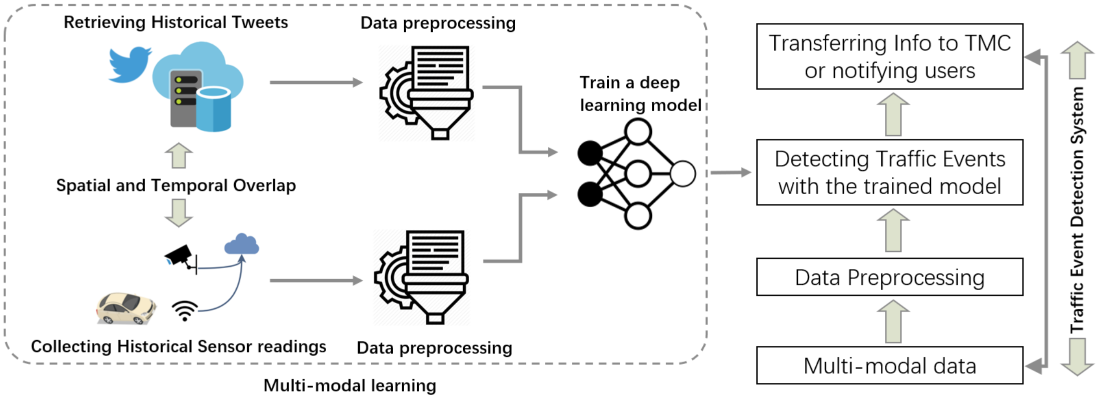

Retrieving Historical Tweets
Data preprocessing
Spatial and Temporal Overlap
Train a deep learning model
Collecting Historical Sensor readings
Data preprocessing
Multi-modal learning
Transferring Info to TMC or notifying users
Detecting Traffic Events with the trained model
Data Preprocessing
Multi-modal data
Traffic Event Detection System

Figure 1: Overview of a traffic event detection system.

search (Wang et al., 2018b; Hou et al., 2018) fuses representations learned from text, visual and audio. Generative Adversarial Network (GAN) (Goodfellow et al., 2014) is one of the most influential models in recent deep learning research. The adversarial learning framework has been adopted in a number of tasks, such as learning representations for realistic image generation (Goodfellow et al., 2014), novelty detection (Sabokrou et al., 2018), and semi-supervised learning (Springenberg, 2015).

Figure 1 depicts the overview of the proposed multi-modal traffic event detection system. It is assumed that when a traffic event occurs, some kind of data characterising the event might be generated at different sources, e.g., pedestrians might post incident information on Twitter, or readings of the traffic sensor might show some different patterns. With the historical data, an event detection model (based on semi-supervised deep learning) can be effectively trained and used to detect future traffic events in real time. Consequently, detail about the event could be transferred to a traffic management centre (TMC) and disseminated to transportation users after verification.

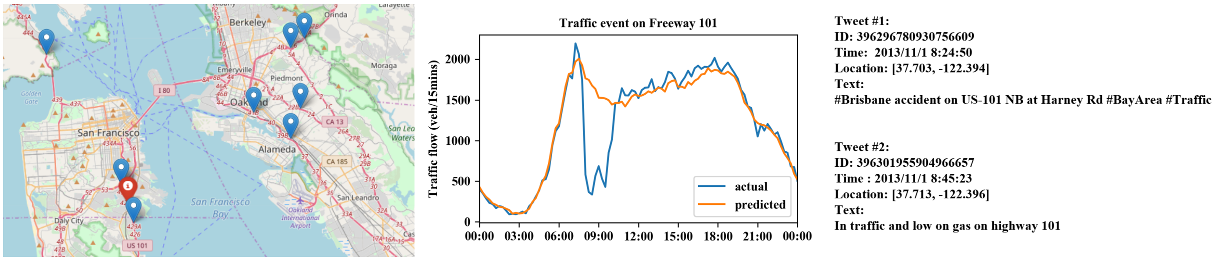
The main contribution of this study is the design and evaluation of a multi-modal Generative Adversarial Network (mmGAN) for the traffic event detection and classification. The proposed network attempts to address the two main limitations of existing studies on integrating data analysis of different modalities and extremely limited amount of labelled data in big data applications. A particularly novel aspect of the network is the employment of semi-supervised learning based on generative adversarial training. To our best knowledge, this is the first work to identify and classify traffic events with both sensor and social media data in a semi-supervised manner. The model has been evaluated on a large, real-world dataset, which contains 20 millions traffic flow readings and 8 millions tweets from the San Francisco Bay Area over a period of 4 months. The results confirmed that mmGAN can effectively learn useful representations characterising the multi-modal data simultaneously.

The rest of the paper is organised as follows. In Section 2, we review some of the representative methods in processing and analysing sensor data and social media textual data in the intelligent transportation domain. In Section 3, we describe in detail the design of the semi-supervised, multi-modal Generative Adversarial Network for traffic event detection and classification, and the algorithm for semi-supervised training. In Section 4, we conduct a number of experiments with the proposed method as well as several baseline models on the same dataset, and discuss the evaluation results. Finally, in Section 5, we conclude the paper and point out some of the future research tasks.

2. Related Work

Traffic events may be caused by many factors, e.g., accidents, traffic hazards, weather conditions, and traffic control. By analysing data collected from the cyber, physical and social worlds, traffic events can be detected and classified. These events are normally reported by transportation authorities, with a possible delay in most of the cases. Figure 2a shows some events reported by the Department of Transportation on November 1st, 2013 (marked in blue and red). Usually, the same event (i.e., the red one in Figure 2a) is signified by data of a single modality, e.g., either sensor data or social media data, as shown in Figure 2b and Figure 2c. However, data from one source of a single modality might be missing, incomplete or even erroneous. Traffic even detection from multi-modal data, e.g., GPS, smartphones, and cameras, has shown impressive performance (Wang et al., 2018b; Hou et al., 2018). In relation to data used in our current work, we discuss the existing work in three categories: sensor data based, social media data based, and multi-modal data based.

Methods using sensor data: With the rapid development in ITSs, the amount of traffic sensor data collected from GPS (Zhang et al., 2015), loop sensors (Lv et al., 2015), smartphones (D'Andrea and Marcelloni, 2017) and cameras (Zhang et al., 2017), is exploding and much of such data has been made publicly available. Sensor observations usually follow a recurring pattern, but may vary abnormally due to traffic incidents, road conditions, social events, and other factors. As illustrated in Figure 2b, the blue curve shows the actual traffic flow, while the orange one depicts the pre-



Tweet #1:
ID: 396296780930756609
Time: 2013/11/1 8:24:50
Location: [37.703, -122.394]
Text:
#Brisbane accident on US-101 NB at Harney Rd #BayArea #Traffic

Tweet #2:
ID: 396301955904966657
Time : 2013/11/1 8:45:23
Location: [37.713, -122.396]
Text:
In traffic and low on gas on highway 101

(a) Traffic events reported by authorities (b) Traffic event detected from sensor observations (c) Traffic event detected from social media data on Twitter

Figure 2: Illustration of the traffic events detected from cyber, physical and social worlds.

dicted traffic flow series. A sudden drop of actual traffic flow at around 9 AM may indicate a potential event and can be detected by anomaly detection applications. Studies in (Zhang et al., 2015) and (D'Andrea and Marcelloni, 2017) develop methods to detect sudden changes in GPS data collected from smartphones and taxi traces to identify incidents, traffic jams, and social events, and further discover when, where and how the events happened.

Methods using social media data: As of 2020, there are around 200 billion tweets posted on Twitter each year. The large amount of social media data, covering nearly everything happening around the world, is easily accessible and has become valuable for research in data mining and knowledge discovery, e.g., sentiment analysis, event detection, and recommendation. In contrast to sensor observation data, social media data has some attractive features, e.g. it covers far more areas and topics, can be collected at low cost, and has high-level semantics understandable to human users. For traffic event detection purpose, millions of geotagged tweets can be acquired from Twitter in real-time and classified using various methods, e.g., Conditional Random Fields (Anantharam et al., 2015), Latent Dirichlet Allocation (Wang et al., 2017), and deep neural networks (Dabiri and Heaslip, 2019).

Methods using multi-modal data: Recent studies (Wang et al., 2018b; Hou et al., 2018) apply deep learning models that fuse representations learned from text, visual and audio, and show superior performance over models based on data of single modality. However, the use of multi-modal data in smart city applications, more specifically, traffic event detection, is still very limited. City CPS data from different sources usually has completely different characteristics. For example, the trustworthiness of the data collected from the social world may be questionable because of social spams. Data from the physical world usually has low-level semantics and may not be always available due to sensor faults or communication failure; furthermore, coverage of the sensor deployment may be limited. In many situations, data from the two worlds can be complementary. By analysing such data together, it is argued that more comprehensive and trustworthy knowledge can potentially be discovered. Previous approaches, i.e., (Pan et al., 2013) and (Anantharam et al., 2016) exploit sensor data for traffic anomalies detection, then search social media data with the detected time and location, and further describe or explain the anomalies with the social media textual data. However, these studies do not consider and process the data of different modalities simultaneously. In a sense, they have not fully exploited the potential of the complementary data. Their limitations are similar to those that process and analyse data of individual modalities. In this paper, we address this issue by designing a multi-modal feature learning component which processes both sensor and social media data simultaneously.

3. multi-modal Generative Adversarial Network (mmGAN) Architecture

The overall architecture of the proposed multi-modal Generative Adversarial Network (mmGAN) is shown in Figure 3. In this architecture, the multi-modal feature learning component is used to encode the input data into numerical vectors and transform the data of different modalities into representations which can be simultaneously processed by one network. The output from each encoders is concatenated to form a multi-modal feature representation for the generative adversarial learning. The semi-supervised Generative Adversarial learning process takes as input the data of multiple modalities and attempts to not only discriminate if the the data is real or generated, but also classify it. It aims to exploit the complementary sensor and social media data for better traffic event detection and classification, with limited amount of labelled data and large amount of unlabelled one.

3.1. Multi-modal Feature Extraction

Our current study considers traffic related data of two different modalities, i.e., sensor data which is usually represented as time series, and social media tweets which are represented as short texts. The multi-modal feature learning architecture transforms different data into a unified multi-modal feature representation as shown in Figure 3. There are two types of encoders: the Sensor Data Encoder component is for sensor input processing (shown in Figure 4) and Social Data Encoder for social media text (shown in Figure 5). The two deep network components extract features from the sensor time-series and twitter messages, respectively. The extracted features are concatenated to form

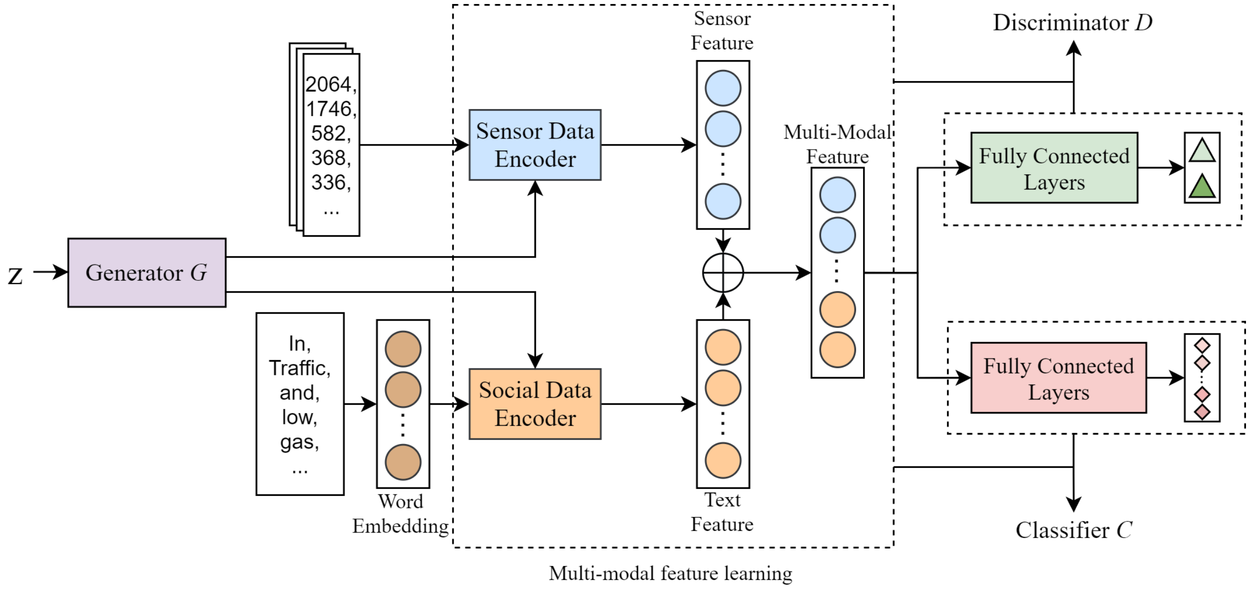


Figure 3: Multi-modal feature learning from both sensor time series and text embeddings

one multi-modal feature representation, which is used in the multi-modal Generative Adversarial Network for detecting and classifying traffic events.

3.1.1. *Sensor Data Encoder*

To extract features from time-series data, we use the Recurrent Neural Network (RNN) as the core module. A RNN contains directed links among neurons, which makes it especially suitable to process data modelled as temporal sequences, $X = (X_1, X_2, ..., X_T)$. At each time step t, the hidden state h_t of the RNN is updated by $h_t = f(h_{t-1}, x_t)$, where f is a non-linear function. We select the LSTM unit in this study which can solve the exploding and vanishing gradient problems of vanilla RNNs. A standard LSTM (Hochreiter and Schmidhuber, 1997) updates the hidden state iteratively with Equation 1:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ C_t &= f_t * C_{t-1} + i_t * tanh(W_C x_t + U_C h_{t-1} + b_C) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ h_t &= o_t * tanh(C_t) \end{aligned} \tag{1}$$

where the output of the forget gate, input gate, and output gate are denoted as f_t, i_t and o_t, respectively. C_t denotes the cell state and h_t denotes the hidden state. The weight W, bias b and sigmoid functions σ are utilised to build connections among input, hidden and output layer.

In Figure 4, two RNN layers are used in the Sensor Data Encoder to extract representations. As traffic sensor observation may vary abnormally due to traffic events, the first RNN layer is pre-trained and aims to predict traffic flow sequences given historical observations. Potential traffic events are represented by the difference between the actual sensor reading and predicted values (referred to as residuals). The calculated residual values during traffic event usually should be much larger than the one at normal period, as shown in Figure 2b. The residual values are the input to the second RNN layer, which aims to extract the representation for the potential events.

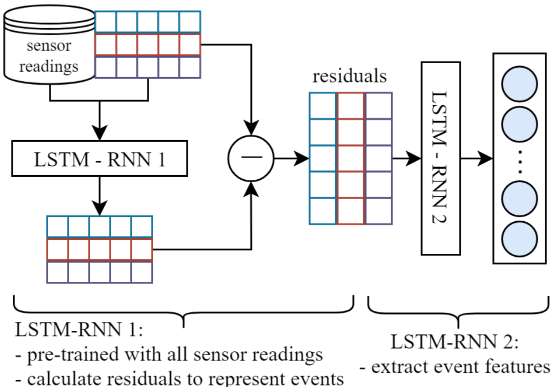


Figure 4: Sensor Data Encoder Architecture

3.1.2. *Social Data Encoder*

The Social Data Encoder component attempts to extract an effective representation for the short social media texts. The input to the Encoder is a sequence of words in a tweet, each of which is represented as a word embedding vector. They are initialised with the word embedding pre-trained on 400 million twitter posts (Godin et al., 2015). A tweet with n words can be represented as $S_{1:n} = (S_1^d, S_2^d, ..., S_n^d)$, where d is the dimension of the embedding vector.

The architecture of encoder is shown in Figure 5. The way that it extracts textual representations from tweets is similar to the one proposed in (Kim, 2014). It consists of a convolutional layer and a max pooling layer. In the convolutional layer, a convolution filter has a size of $h \times d$, where h is the window size and d is the width of filter equal to the word-vector dimension. Sliding the filter across the matrix $S_{1:n}$ produces a feature map s^j with size $(n-h+1)$ which is represented as $s^j = [s_1, s_2, ...s_i, ..., s_{n-h+1}]$, where s_i is calculated with Equation 2.

$$s_i = ReLU(W_c \cdot S_{i:i+h-1}) \tag{2}$$

where $ReLU$ is an activation function, W_c represents the weights of filter, $S_{i:i+h-1}$ represents the contiguous h word embedding vectors and $(\cdot)$ is the dot product between weights W_c and word vectors $S_{i:i+h-1}$. As shown in Figure 5, the coloured dashed lines in the convolutional layer represents this convolutional learning process, where different colours denote different filters.

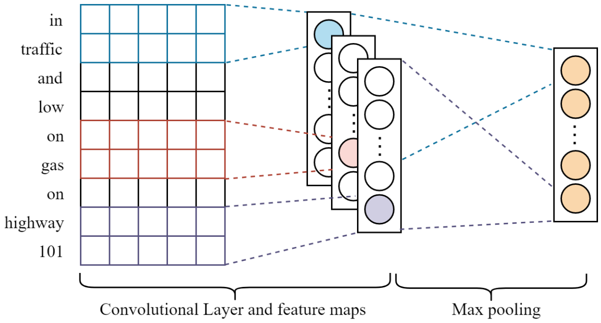


Figure 5: Social Data Encoder architecture

With K different filters, K feature maps $s = [s^1, s^2, ..., s^K]$ are generated. We apply a max-pooling operation to each feature map with Equation 3.

$$m = [\delta(s^1), \delta(s^2), ..., \delta(s^K)] \tag{3}$$

where $\delta(s^j)$ denotes the max-pooling process, which selects the maximum value from the feature map s^j. Intuitively, the convolution operation extracts local features into higher level representations in the feature maps, and the max-pooling extracts the distinguishing aspects of each feature map while reducing the output dimension. The final social data representation m with size K is then concatenated with sensor data representation for further multi-modal classification.

3.2. Semi-supervised Generative Adversarial Learning

Due to the fact that only limited amount of labelled data is available in any big data applications, it is not appropriate to use standard deep learning methods for supervised tasks. We noticed in our experiments that results generated using either the Sensor Data Encoder, Social Data Encoder or Multi-modal feature component is unsatisfactory. To this end, we extend the model to a semi-supervised architecture based on the Generative Adversarial Network (GAN) (Goodfellow et al., 2014). The original GAN sets up an adversarial game between a discriminator and a generator. The goal of the discriminator is to distinguish whether a sample is drawn from the true data or generated by the generator. On the contrary, the generator is optimised to produce samples that are not distinguishable by the discriminator. In this way, the generator and discriminator compete with each other to boost their performance in a seamless manner.

A standard supervised model, e.g., classifiers, can be extended to a semi-supervised one by adding samples from the generator G in GAN to the dataset and labelling them with a new class, "generated", denoted as $y = K + 1$ (Springenberg, 2015; Salimans et al., 2016; Dai et al., 2017; Zheng et al., 2017). Inspired by the idea, we design a multi-modal, semi-supervised adversarial training architecture for traffic event detection and classification.

As shown in Figure 3, the architecture consists of three components: a Generator G, a Discriminator D, and a Classifier C. G and D in mmGAN are trained with conflicting objectives. G takes in a noise vector z and produces traffic data of two modalities: sensor observations and social media texts (both in the form of numerical vectors). D takes in multi-modal feature vector and predicts if it is a sample from the real data or G. G is trained to maximise the probability that D makes a mistake, while D is trained to minimise the probability that it makes a wrong prediction. Through this adversarial training process, features that could distinguish real samples from the generated ones are learned in an unsupervised way. Meanwhile, the multi-modal feature component could learn representations from a large amount of unlabelled data through this adversarial training process.

C is a standard multi-class classifier that also takes in multi-modal feature vector and attempts to predict a correct label for an input. As the multi-modal feature learning component is shared by both D and C, the three components can be jointly optimised. The representations learned from the unlabelled data could also help improve the performance of C. The loss function for training the generator L_g, discriminator L_d and classifier L_c are shown in Equation 4, Equation 5, and Equation 6 respectively.

$$L_g = -\mathbb{E}_z \log D(G(z)) \tag{4}$$

$$\begin{aligned} L_d = & -\mathbb{E}_{x_s, x_t \sim p_{data}(x_s, x_t)} \log D(x_s, x_t) \\ & -\mathbb{E}_z \log(1 - D(G(z))) \end{aligned} \tag{5}$$

$$L_c = -\mathbb{E}_{x_s, x_t, y \sim p_{data}(x_s, x_t, y)} \log C(x_s, x_t) \tag{6}$$

where x_s and x_t represents sensor data input and twitter word embedding input, respectively. The detailed training process of the proposed mmGAN is summarised in Algorithm 1.

4. Experiments and Evaluation

We have conducted extensive experiments using a large, real world, multi-modal dataset, which was prepared by in-

Algorithm 1 mmGAN Training Algorithm

```
Input: unlabelled multi-modal input $(x_s, x_t)$, and labelled multi-modal input $(x_s, x_t, y)$;
for the number of training iterations do
    Draw $m$ noise samples;
    Draw $m$ samples from unlabelled multi-modal input $(x_s, x_t)$;
    Perform gradient descent on the parameters of $D$ according to Eq. 5 on the combined mini-batch of size $2m$;
    Draw $m$ noise samples;
    Perform gradient descent on the parameters of $G$ according to Eq. 4;
    Draw $m$ samples from labelled multi-modal input $(x_s, x_t, y)$;
    Perform gradient descent on the parameters of $C$ according to Eq. 6;
end for
```

terlinking two large datasets that have been widely used in existing research and a specific dataset constructed by ourselves. Performance of the proposed method was evaluated and compared to a number of state-of-the-art supervised learning methods.

4.1. Dataset

The Caltrans Performance Measurement System (PeMS) (Caltrans, 2019) provides large amount of traffic sensor data that has been widely used by the research communities. The data is collected every 30 seconds from many vehicle detector stations that report at 5-minute interval throughout the state of California in the United States. We used the traffic flow data in the San Francisco Bay Area from August 2013 to November 2013 and further aggregated the readings at 15-minute interval. The resulting dataset contains around 20 million traffic flow readings from 1,649 traffic detector stations and data size is around 13 GB.

We reused the geo-tagged twitter dataset published in (Anantharam et al., 2015), in which more than 8 million tweets were collected from August 2013 to November 2013 for the San Francisco Bay Area. The size of the twitter dataset is around 2 GB. Most of the tweets were posted by ordinary people that cover a variety of topics. We also collected geo-tagged tweets that report traffic event information from official accounts, i.e., @TotalTrafficSF, from August 2013 to November 2013 with the twitter API.

To create the multi-modal traffic dataset, we first filtered traffic-related tweets with a list of keywords, e.g. "traffic, block, delay, highway, freeway, accident, incident, construction". We matched the tweets with the sensor data that has temporal (two hours) and spatial (one kilometer) overlapping, therefore creating 2,227 pairs of multi-modality samples. We used the (California Highway Patrol) CHP incident dataset collected from the PeMS (Caltrans, 2019) to assist the labelling process. The CHP incident dataset contains the detailed time, location, duration and incident types (e.g., traffic hazard, traffic collision, etc.), which we used to label the multi-modal instances based on spatial and temporal overlaps. The remaining instances that are not covered in the incident dataset are manually labelled. Our task is to consider both the sensor and tweet data simultaneously and categorise it into one of the two classes: (1) **Traffic event**, representing a non-recurring event that generates an abnormal change in traffic and transportation capacity. The examples of non-recurring events include traffic crashes, disabled vehicles, road construction, vehicle fire, etc. The current work is to inform users and agencies the occurrence of an ongoing traffic event if there is any. We will consider the case of multi-class classification in the future work, which would provide users more intuitive information with specific event types; and (2) **Traffic information (non-traffic event)**, reporting daily traffic conditions, past traffic events, new traffic rules, traffic advisory, and any other information on transport infrastructures. The number of traffic event and the number of traffic information for both training and test sets are reported in Table 1.

Table 1
Distributions of the multi-modal datasets

	Training Dataset	Test Dataset	Total
Traffic Event	1390	155	1545
Traffic Info	614	68	682
total	2004	223	2227

4.2. Setup

Each input sample to the model consists of a sensor observation sequence and a twitter message (in the form of a sequence of word embeddings). As there may be multiple sensors reporting the same event, we built an input block for sensor data with a block size of 10. The dimension of the sensor input shape is 16×10, where 16 is the number of time steps in 4 hours. For social media text input, we represented each word with a word embedding of 400 dimension. Most of the tweets contain less than 15 words, so we only considered the first 15 words in each tweet. The dimension of a twitter input is 15×400.

We performed stratified 10-fold cross-validation and keep each partition containing roughly the same proportions of traffic event instances and traffic information instances. The model is trained with 90% of the data and tested with the rest 10% of the data. We performed a grid search to determine the best parameters for the proposed mmGAN: in the Social Data Encoder, the window size of filter was set to 2, and the dimension of the hidden units in both Social Data Encoder and Sensor Data Encoder was set to 32. For the two fully connected layers in Discriminator D and Classifier C, the hidden size was set to 32. The number of batch size was 64; the dropout rate was set to 0.5; the Adam optimiser with early stopping was used to avoid overfitting.

Table 2
Performance Comparison of Different Models

Data	Sensor Data Only		Social Data Only		Multi-modal Data		
Models	SVM	Sensor Data Encoder	SVM	Social Data Encoder	RNN	CNN	mmGAN
Accuracy	60.83 ± 5.34	68.23 ± 2.50	76.10 ± 3.75	82.24 ± 2.09	84.17 ± 2.08	83.45 ± 1.86	**87.17 ± 3.63**
Precision	59.53 ± 4.38	67.93 ± 7.23	76.76 ± 3.35	83.19 ± 2.51	84.30 ± 1.74	83.21 ± 2.19	**87.40 ± 3.65**
Recall	60.83 ± 5.34	68.28 ± 2.50	76.10 ± 3.75	82.24 ± 2.09	84.17 ± 2.08	83.34 ± 1.86	**87.17 ± 3.63**
F1	59.87 ± 4.55	64.36 ± 9.53	76.30 ± 3.51	82.37 ± 2.18	84.05 ± 1.92	83.21 ± 2.14	**87.16 ± 3.53**
AUC	60.06 ± 7.82	64.94 ± 7.62	82.31 ± 3.07	90.87 ± 1.43	91.43 ± 2.69	91.03 ± 2.31	**93.44 ± 2.06**

4.2.1. Baseline Models

Data of single modality (i.e., either sensor data or social media data) can also be used to discover traffic events. We re-implemented and tested the following two baseline models with only data of single modality: (1) **Support Vector Machine (SVM)** is a popular kernel method for supervised learning tasks and has been widely used to process sensor data and social media data, e.g., time series prediction (Vanajakshi and Rilett, 2007) and twitter classification (Pereira et al., 2017). We implemented two SVM models using the sensor time series and social text embeddings, respectively, to detect traffic events; and (2) **Sensor Data Encoder** and **Social Data Encoder** have been explained in Section 3, and shown in Figure 4 and Figure 5, respectively. They were used separately to extract features from sensor data and social media data. To detect and classify traffic events, a fully connected layer with hidden size of 32 and a sigmoid output layer were added at the top of the two models.

We also re-implemented and tested the following three baseline models with data multiple modalities for performance comparison: (1) **Multi-modal Network (MMN)** was implemented as a classifier and its architecture is similar to the component C in the mmGAN architecture; (2) **RNN** has been primarily used to to extract features from data with strong temporal characteristics, e.g., sensor data (Tian and Pan, 2015) and social media data (Dabiri and Heaslip, 2019). In our implementation, after using RNN to process sensor data and social media data separately, the extracted feature vectors were concatenated and fed into a fully connected layer for prediction; and (3) **CNN** has also been widely employed in many different types of supervised tasks, e.g., image classification (He et al., 2016), and natural language processing (Gehring et al., 2017). It has also been successfully used to extract features from sensor data (Wang et al., 2018a) and social media data (Dabiri and Heaslip, 2019). We used CNN to extract features from sensor and social media data separately, and concatenated the feature vectors for prediction.

4.3. Evaluation

We used the standard evaluation metrics for assessing classification performance, i.e., accuracy, weighted average recall, precision, $F1$ and AUC score. We report the mean and and the 95% confidence interval of the testing results on models trained with 10-fold cross-validation. The evaluation results of the proposed mmGAN and the baseline models are shown in Table 2.

The most notable observation is that the overall performance of those models that process and analyse multi-modal data simultaneously is better than those with only data of a single modality in all metrics. This observation obviously confirmed our expectation that exploiting complementary data of different modalities does enable us to extract trustworthy knowledge and improve classification performance. Among the three models that process multi-modal data, mmGAN outperformed both CNN and RNN with 3% improvement in accuracy, 3.1% in precision, 3% in recall, 3.11% in $F1$ and 2.01% in AUC. This showed the effectiveness of the multi-modal learning and the adversarial training: the sensor and social data encoder learn representations from both types of data; the generator and the discriminator compete with each other and improve each others' performance at the same time. As a consequence, the classifier making use of the shared component in the discriminator is able to improve its own performance with the multi-modal data, even though its differs greatly in the level of granularities and semantic meaning.

Table 2 shows that models (i.e., SVM, Social Data Encoder, and Sensor Data Encoder) which process and analyse single modality data can also detect and classify events successfully with certain degree. In general, as sensor data contains much noise and many missing values, e.g., no nearby sensors to report such traffic event, models using the sensor data produced the worst performance among all the methods. On the contrary, social media data contains more obvious features, e.g., traffic events related keywords, which helped extract more informative representations and generate better results compared with processing sensor data. With single modality data, the proposed Sensor Data Encoder and Social Data Encoder outperformed the SVM in terms of all metrics, which showed that the two models could extract reasonably good representations from the sensor and text data, respectively.

To evaluate the performance of the mmGAN with a limited amount of labelled data, we compared mmGAN with the Multi-modal Network (MMN). MMN is a classifier and its architecture is identical to the component C in the mmGAN architecture. Their performance in terms of accuracy with different amount of labelled data is plotted in Figure 6. It can be seen that the proposed mmGAN outperformed

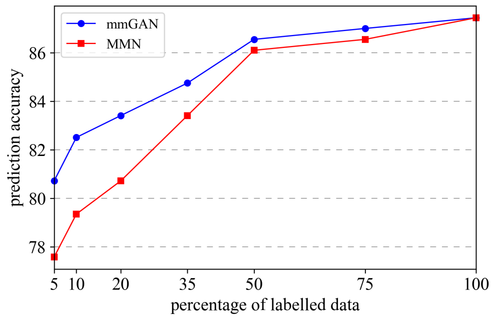


Figure 6: Classification accuracy of mmGAN and MMN with different amount of labelled data.

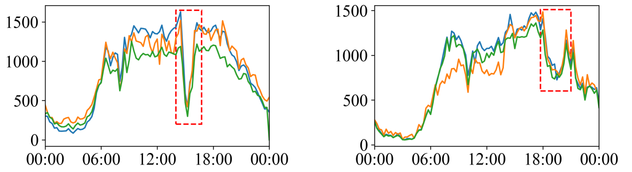
MMN when we shrank the size of the training data. When the size of the labelled data was very small, e.g., $5\% - 30\%$ of the whole labelled data, mmGAN notably outperformed MMN. This confirmed that the Discriminator D could learn better representations from both large amount of unlabelled and very limited amount of labelled data through the generative adversarial training. Ultimately, it contributed to improve the performance of Classifier C. It can also be seen that when more and more amount of labelled data was used, their performance tended to converge. As the number of traffic related tweets that can be matched with sensor readings is limited, we select a portion of the paired data as labelled data and treat the rest part as unlabelled data. We will consider using the additional unlabelled data (13GB sensor data and 2GB tweet data) of each modality separately in an unsupervised manner to further improve the results in the future work.

4.4. Case Study

We provided 4 examples extracted from the dataset in Figure 7 and 8. They were misclassified by models using a single modality data, but correctly classified with mmGAN using multi-modal data. We show the sensor readings and social media content in four subfigures. In each subfigure, the coloured lines show the patterns of real sensor readings, where x-axis represents time of day and y-axis denotes traffic flow in 15 minutes. The colours of lines represent different sensors close to (within 1 kilometer) a particular event, and the number of sensors reported in each event could be different. The corresponding tweet messages are shown below.

Figure 7 shows two traffic events that were successfully detected by mmGAN but were missed by single modality model. The tweet messages did not provide enough evidence to identify whether it was a traffic event, so the Social Encoder model wrongly classified the two cases as non traffic events. By adding the sensor data into the proposed mmGAN, the cases were correctly classified as traffic events.

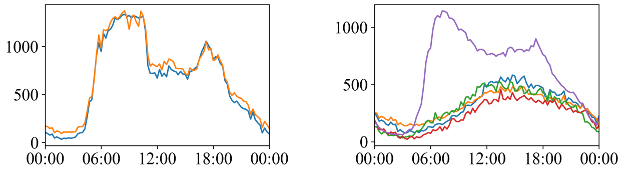
Figure 8 shows two examples misclassified as real-time traffic events by models using only the social media data, but successfully classified by mmGAN. As the two tweets contain traffic event related keywords, e.g. stuck, traffic, and accident, the Social Encoder model categorised them as traffic events. However, the traffic flow data at the same location during the same time period showed normal traffic patterns. By using both type of data, the cases were correctly classified as non-traffic events by mmGAN.



(a) 2013-09-17 15:01:42
Forgot how much I hate traffic on the 101. And why didn't I get a tesla rental at the Avis counter?

(b) 2013-10-02 19:13:00
Sun is going down. Stuck in traffic. Ugh. Looks like we're going to be late to #maroon5.

Figure 7: Two traffic events correctly detected by mmGAN but misclassified by others.



(a) 2013-09-03 19:08:49
Reporter on the San Jose news just said higher employment = more traffic. Now we can get stuck in traffic.

(b) 2013-09-02 08:29:35
SJPD seek driver after double-fatal accident.

Figure 8: Traffic related information detected by mmGAN but misclassified by single modality model

5. Conclusion and Future Work

Modern expert systems and applications usually process and analyse big data from different sources, often in multiple modalities. While the data might be noisy, incomplete or inconsistent, it is complementary to each other and potentially enables more valuable knowledge to be extracted. In addition, it is also impractical to obtain large amount of labelled data in real-world applications. We propose the multimodal Generative Adversarial Network, a semi-supervised, deep learning based model that can process data of multiple modalities in a unified framework, for traffic event detection and classification. The evaluation results clearly showed the advantages of the mmGAN over other models with or without multi-modal data in terms of precision, accuracy and $F1$ in classification. Furthermore, the generative adversarial training process with large amount of unlabelled and limited amount of labelled data could indeed help extract more useful knowledge than other baseline models.

In the current work, we only focused on traffic sensor data and textual data. In the future, we plan to further refine the proposed model so that it can process more types of data for other smart city applications, e.g., GPS traces, image and wearable sensor data. As data from different sources cannot always be matched, an interesting direction is to extend the proposed model to be compatible with both single modality input and multi-modality input. An important problem being considered is to apply the attention mechanisms to model the temporal interplay of multi-modal data and learn better representations from it, rather than simply concatenating representations learned from data of individual modality. Another future work is to extend the current model to support multi-class or even multi-label classification, which would provide users more intuitive knowledge.

Acknowledgment

This research was funded by the Research Development Fund at Xi'an Jiaotong-Liverpool University, contract number RDF-16-01-34. It is also partially supported by the following: National Natural Science Foundation of China under no.61876155; Natural Science Foundation of Jiangsu Province BK20181189; Key Program Special Fund in XJTLU under no. KSF-A-01, KSF-T-06, KSF-E-26, KSF-P-02 and KSF-A-10.

References

Anantharam, P., Barnaghi, P., Thirunarayan, K., Sheth, A., 2015. Extracting city traffic events from social streams. Acm Transactions on Intelligent Systems & Technology 6, 1–27.

Anantharam, P., Thirunarayan, K., Marupudi, S., Sheth, A., Banerjee, T., 2016. Understanding city traffic dynamics utilizing sensor and textual observations, in: Thirtieth AAAI Conference on Artificial Intelligence, pp. 3793–3799.

Caltrans, 2019. Performance measurement system (pems). URL: `http://pems.dot.ca.gov.`. [Online].

Dabiri, S., Heaslip, K., 2019. Developing a twitter-based traffic event detection model using deep learning architectures. Expert Systems with Applications 118, 425–439.

Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R.R., 2017. Good semi-supervised learning that requires a bad gan, in: Advances in Neural Information Processing Systems, pp. 6510–6520.

D'Andrea, E., Marcelloni, F., 2017. Detection of traffic congestion and incidents from gps trace analysis. Expert Systems with Applications 73, 43–56.

Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N., 2017. Convolutional sequence to sequence learning, in: ICML, PMLR. pp. 1243–1252.

Godin, F., Vandersmissen, B., De Neve, W., Van de Walle, R., 2015. Multimedia lab@ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations, in: Proceedings of the workshop on noisy user-generated text, pp. 146–153.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

Gu, Y., Qian, Z., Chen, F., 2016. From twitter to detector: real-time traffic incident detection using social media data. Transportation Research Part C 67, 321–342.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Hou, J.C., Wang, S.S., Lai, Y.H., Tsao, Y., Chang, H.W., Wang, H.M., 2018. Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence 2, 117–128.

Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 .

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y., 2015. Traffic flow prediction with big data: a deep learning approach. IEEE Transactions on Intelligent Transportation Systems 16, 865–873.

Pan, B., Zheng, Y., Wilkie, D., Shahabi, C., 2013. Crowd sensing of traffic anomalies based on human mobility and social media, in: Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 344–353.

Pereira, J., Pasquali, A., Saleiro, P., Rossetti, R., 2017. Transportation in social media: an automatic classifier for travel-related tweets, in: EPIA Conference on Artificial Intelligence, Springer. pp. 355–366.

Sabokrou, M., Khalooei, M., Fathy, M., Adeli, E., 2018. Adversarially learned one-class classifier for novelty detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3379–3388.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans, in: Advances in neural information processing systems, pp. 2234–2242.

Song, X., Kanasugi, H., Shibasaki, R., 2016. DeepTransport: prediction and simulation of human mobility and transportation mode at a citywide level., in: IJCAI, pp. 2618–2624.

Springenberg, J.T., 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390 .

Tian, Y., Pan, L., 2015. Predicting short-term traffic flow by long short-term memory recurrent neural network, in: IEEE International Conference on Smart City/socialcom/sustaincom, pp. 153–158.

Valipour, S., Siam, M., Stroulia, E., Jagersand, M., 2016. Parking-stall vacancy indicator system, based on deep convolutional neural networks, in: Internet of Things, pp. 655–660.

Vanajakshi, L., Rilett, L.R., 2007. Support vector machine technique for the short term prediction of travel time, in: 2007 IEEE Intelligent Vehicles Symposium, IEEE. pp. 600–605.

Wang, D., Al-Rubaie, A., Clarke, S.S., Davies, J., 2017. Real-time traffic event detection from social media. ACM Transactions on Internet Technology (TOIT) 18, 9.

Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L., 2018a. Deep learning for sensor-based activity recognition: A Survey. Pattern Recognition Letters .

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J., 2018b. Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM. pp. 849–857.

Zhang, S., Wu, G., Costeira, J.P., Moura, J.M., 2017. Understanding traffic density from large-scale web camera data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5898–5907.

Zhang, W., Qi, G., Pan, G., Lu, H., Li, S., Wu, Z., 2015. City-scale social event detection and evaluation with taxi traces. ACM Transactions on Intelligent Systems and Technology (TIST) 6, 40.

Zhang, Z., He, Q., Jing, G., Ming, N., 2018. A deep learning approach for detecting traffic accidents from social media data. Transportation Research Part C Emerging Technologies 86, 580–596.

Zheng, Z., Zheng, L., Yang, Y., 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3754–3762.