

# CASTNet: Community-Attentive Spatio-Temporal Networks for Opioid Overdose Forecasting

Ali Mert Ertugrul<sup>1,2</sup>, Yu-Ru Lin<sup>1</sup>(✉), and Tugba Taskaya-Temizel<sup>2</sup>

<sup>1</sup> University of Pittsburgh, Pittsburgh PA, USA {ertugrul,yurulin}@pitt.edu

<sup>2</sup> Middle East Technical University, Ankara, Turkey ttemizel@metu.edu.tr

**Abstract.** Opioid overdose is a growing public health crisis in the United States. This crisis, recognized as “opioid epidemic,” has widespread societal consequences including the degradation of health, and the increase in crime rates and family problems. To improve the overdose surveillance and to identify the areas in need of prevention effort, in this work, we focus on forecasting opioid overdose using real-time crime dynamics. Previous work identified various types of links between opioid use and criminal activities, such as financial motives and common causes. Motivated by these observations, we propose a novel spatio-temporal predictive model for opioid overdose forecasting by leveraging the spatio-temporal patterns of crime incidents. Our proposed model incorporates multi-head attentional networks to learn different representation subspaces of features. Such deep learning architecture, called “community-attentive” networks, allows the prediction for a given location to be optimized by a mixture of groups (i.e., *communities*) of regions. In addition, our proposed model allows for interpreting what features, from what communities, have more contributions to predicting local incidents as well as how these communities are captured through forecasting. Our results on two real-world overdose datasets indicate that our model achieves superior forecasting performance and provides meaningful interpretations in terms of spatio-temporal relationships between the dynamics of crime and that of opioid overdose.

**Keywords:** Forecasting opioid overdose · Spatio-temporal networks · Multi-head attentional networks · Crime dynamics.

## 1 Introduction

Opioid use disorders and overdose rates in the United States have increased at an alarming rate since the past decade [25]. Overdose deaths have risen since the 1990s, and the number of heroin overdose deaths has risen sharply since 2010 [21]. Mortalities caused by drug overdose are the main cause of injury-related death in the U.S. [3]. The growth rate of opioid overdose together with the number of impacted individuals in the U.S., has led many to classify this as an “opioid epidemic” [16]. Enhanced understanding of the dynamics of the overdose epidemic may help policy-makers to develop more effective epidemic prevention mechanisms and control strategies [13].

The opioid epidemic is a complex social phenomenon involving and interacting with various social, spatial and temporal factors [2]. Highlighting the links between opioid use and various factors has drawn significant attention. Among those, studies

have identified relationships between opioid use and crime incidences, including cause (that opioid use leads to criminal activities [1]), effect (that involvement in criminal behavior leads to drug use [10]), and common causes (that crime and drug tend to co-occur [23]). Crime occurrences also have non-trivial spatio-temporal characteristics – for example, routine activity theory suggested that crimes may exhibit spatio-temporal lags as the *likely offenders* of one place may reach *suitable targets* in other places. Therefore, how to unveil the complicated relationship between opioid use and crime incidences is challenging.

In seeking data for investigation, detailed assessments of opioid use disorders and overdose growth require systematically collected well-resolved spatio-temporal data [9]. However, the amount of systematically monitored data either at a regional or local level in the U.S. is very limited. In addition, there is no common reporting mechanism for incidents. For instance, the incident categories and the organization of categories vary significantly across the databases. On the other hand, crime data is meticulously collected, organized and stored, at a finer-grained level. Given the plausible relationship between the crime dynamics and opioid use as well as the availability of real-time crime data for various locations, in this study, we seek to explore the capability of forecasting opioid overdose using real-time crime data.

Recent works in predictive modeling has shown significant improvement in spatio-temporal event forecasting and time series prediction [20, 26]. However, these studies suffer from two main concerns. First, most of them overlook the complex interactions between local (observed from within a region) and global (observed from all regions) activities across time and space. Only a few have paid attention to this problem, yet they model the global activities as a single universal representation [6, 5], which is either irrespective of the event location or is reweighed based on a pre-defined fixed proximity matrix [17]. In other words, none of the existing works learns to differentiate the pairwise activity relationships between a particular event location and other locations. Second, most of the spatio-temporal forecasting studies mainly focus on prediction performance and lack interpretability to uncover the underlying spatio-temporal characteristics of the activities, such as (1) what local and global activity features are more predictive for the subsequent events? (2) what are the locations that have more salient contribution to predicting opioid overdose with respect to the target location? Inspired by the idea of multi-head attentional networks [24], in this work, we propose a novel deep learning architecture, called “CASTNet,” for opioid overdose forecasting using spatio-temporal characteristics of crime incidents, which seeks to address the aforementioned problems. Assuming that different locations could share similar dynamics, our approach aims to learn different representation subspaces of cross-regional dynamics, where each subspace involves a set of locations called “community” that share similar behaviors. The proposed architecture is called “community-attentive” as it allows the prediction for a given location to be individually optimized by the features contributed by a mixture of communities. Specifically, combining the features of the given target location and features from the communities (referred to as local and global dynamics), the model learns to forecast the number of opioid overdoses in the target location. Meanwhile, by leveraging a Lasso regularization [22] and hierarchical attention mechanism, our method allows for interpreting what local and global features are more

predictive, what communities contribute more to predicting incidences at a location, and what locations contribute more to each community.

Overall, our contributions include: (1) *A community-attentive spatio-temporal network*: We propose a multi-head attention based architecture that learns different representation subspaces of global dynamics (communities) to effectively forecast the opioid overdoses for different target locations. (2) *Interpretability in hierarchical attention and features*: First, CASTNet incorporates a hierarchical attention mechanism which allows for interpreting community memberships (which locations form the communities), community contributions for forecasting local incidents and informative time steps in both local and global for the prediction. Second, CASTNet incorporates Group Lasso (GL) [22] to select informative features which succinctly capture what activity types at both local- and global-level are more associated with the future opioid overdoses. (3) *Extensive experiments*: We performed extensive experiments using real-world datasets from City of Cincinnati and City of Chicago. The results indicate a significant improvement in forecasting performance with greater interpretability compared to several baselines and state-of-the-art methods.

## 2 Related Work

The existing works have investigated the links between opioid use and various social phenomena as well as contextual factors including crime and economic stressors. Among them, Hammersley et al. [10] stated that opportunities for drug use increase with involvement in criminal behavior. The people dependent on opiates are disproportionately involved in criminal activities [1] especially for the crimes committed for financial gain [19]. Seddon et al. [23] revealed that crime and drug use share common set of causes and they co-occur together. Most of the existing works studying the relationship between opioid use and social phenomena have employed basic statistical analysis, and focused on current situation and trends rather than predicting/forecasting overdose. Moreover, these studies overlooked the interactions among spatio-temporal dynamics of the locations. Among the studies predicting/forecasting opioid overdose, regression-based approaches have been applied in individual-level [8] and state-level [14]. Also, a neural network-based approach has been proposed [6] to forecast heroin overdose from crime data, which identifies the predictive hot-spots. However, the effect of these hot-spots is universal and irrespective of event locations on the prediction.

Furthermore, there have been studies that utilized spatial and temporal dependencies for event forecasting and time series prediction. With the success of neural network-based models, several studies employed neural models to forecast/detect events related to anomalies [4], crime [12] and social movements [5]. Additionally, several studies utilized deep neural models for times series prediction. Among them, Ghaderi et al. [7] proposed a recurrent neural network (RNN) based model to forecast wind speeds. Qin et al. [20] presented a dual-stage attention-based RNN model to make time series prediction. Similarly, Liang et al. [17] proposed multi-level attention networks for geo-sensory time series prediction. A few of the studies considered the complex relationships between local and global activities, yet they modeled the global activities as a universal representation, which either does not change from event location to location or is adjusted by a pre-defined fixed proximity matrix. Most of these works simply employed a single temporal model to model various local and global spatio-temporal

activities, which is insufficient to capture the complex spatio-temporal patterns at both local and global levels. Moreover, existing methods primarily focus on forecasting performance, yet they provide no or limited interpretability capability to unveil the underlying spatio-temporal characteristics of the local and global activities.

### 3 Method

#### 3.1 Problem Definition

Suppose there are  $L$  locations-of-interest (e.g. neighborhoods, districts) and each location  $l$  can be represented as a collection of its static and dynamic features. While the static features (e.g. demographics, economical indicators) remain same or change slowly over a longer period of time, the dynamic features are the updates for each time interval  $t$  (e.g. day, week). Let  $X_l^{stat}$  be the static features of location  $l$ , and  $X_{t,l}^{dyn}$  the set of dynamic features for location  $l$  at time  $t$ . We are also given a continuous variable  $y_{t^*,l} \in \mathbb{N}$  that indicates the number of opioid overdose incidents (e.g. emergency medical services (EMS) calls, deaths) at location  $l$  at future time  $t^*$ . The collection of dynamic features from all locations-of-interest within an observing *time window* with size  $w$  up to time  $t$  can be represented as  $\mathcal{X}_{t-w+1:t}^{dyn} = \{\mathcal{X}_{t-w+1}^{dyn}, \dots, \mathcal{X}_t^{dyn}\}$ , where  $\mathcal{X}_t^{dyn} = \{X_{t,1}^{dyn}, \dots, X_{t,L}^{dyn}\}$ .

Our goal is to predict the number of future opioid overdose incidents  $y_{t^*,l}$  at specific location  $l$  at a future time  $t^* = t + \tau$ , where  $\tau$  is called the *lead time* for forecasting. The forecasting is based on the static and dynamic features of the target location itself, as well as the dynamic features in the environment (from all locations-of-interest). Therefore, the forecasting problem can be formulated as learning a function  $f(X_d^{stat}, \mathcal{X}_{t-w+1:t}^{dyn}) \rightarrow y_{t^*,d}$  that maps the static and dynamic features to the number of opioid overdose incidents at the future time  $t^*$  at a *target* location  $d$ .

To facilitate spatio-temporal interpretation of the forecasting, we seek to develop a model that can differentiate contribution of the features, the locality (local features vs. global features) and the importance of latent communities when contributing to the prediction of other locations. Therefore, we further organize the dynamic features  $\mathcal{X}_{t-w+1:t}^{dyn}$  into two sets: the *local* features,  $\{X_{t-w+1,d}^{dyn}, \dots, X_{t,d}^{dyn}\}$  represent dynamic features for the target location  $d$ , and the *global* features,  $\{X_{t-w+1,l}^{dyn}, \dots, X_{t,l}^{dyn}\}$  for  $l \in \{1, 2, \dots, L\}$ , contain the sequences of dynamic features for all locations of interest.

#### 3.2 Architecture

In this work, we propose an interpretable, community-attentive, spatio-temporal predictive model, named CASTNet. As shown in Fig. 1, our proposed architecture consists of three primary components, local component (Fig. 1a), static component (Fig. 1b) and global component (Fig. 1c). The global component is designed to model the global contribution of dynamic features for all-locations-of-interest by learning different representation subspaces of global dynamics, and to output target location-specific global contribution. On the other hand, the local component is designed to model the contribution of local dynamic features for the target location. Finally, the static component models location-specific static information about the target location.

**Global Component.** This component produces the target location-specific global contribution (from all locations) to forecast the number of incidents at the target location  $d$  at future time  $t^*$ . It consists of  $K$  number of community blocks, where each community

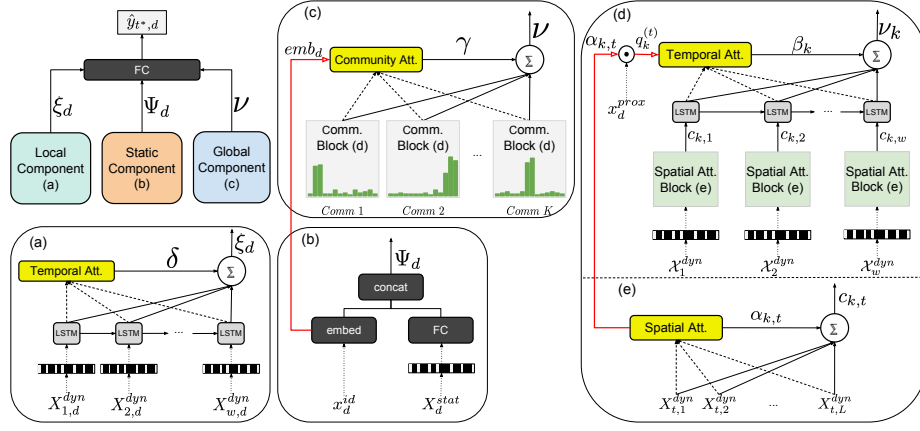


Fig. 1: **Overview of our proposed CASTNet architecture.** The local component (a) models local dynamics of the locations, and the static component (b) models the static features. The global component (c) summarizes different representation subspaces (i.e. communities) of global dynamics, learned by community blocks (d), by querying these multi-subspace representations through the embedding of the target location ( $emb_d$ ). Spatial Att. Block (e) reweights the global dynamics of locations. Checkered rectangles on top of the inputs represent GL regularization. Red arrows indicate the queries for the corresponding attentions. “FC”: fully-connected layer; “embed”: embedding layer.

block learns a different representation subspace of the global dynamic features, which is inspired by the idea of multi-head attention [24]. A community block (Fig. 1d) models the global dynamic features through a hierarchical attention network which consists of a spatial attention block (Fig. 1e), a recurrent unit and a temporal attention. For the sake of clarity, we explain the internal mechanism of global component in a bottom-up manner by following the order (Fig. 1e  $\rightarrow$  1d  $\rightarrow$  1c):

**Spatial Attention Block** is used to reweight the contribution of dynamic features of each location  $i$  at time  $t$ . More specifically, the attention weight,  $\alpha_{k,t}^{(i)}$ , represents the contribution of the location  $i$  at time  $t$  to the community  $k$ . Since higher spatial attention weight for a location indicates the involvement of its dynamic features in this community, we call this community membership.  $c_{k,t}$  is the context vector, which summarizes the aggregated contribution of all locations as follows:

$$e_{k,t} = (v_k^{sp})^\top \tanh(W_k^{sp} \mathcal{X}_t^{dyn} + b_k^{sp}) \quad (1)$$

$$\alpha_{k,t}^{(i)} = \frac{\exp(e_{k,t}^{(i)})}{\sum_{l=1}^L \exp(e_{k,t}^{(l)})}; \quad c_{k,t} = \sum_{l=1}^L \alpha_{k,t}^{(l)} X_{t,l}^{dyn} \quad (2)$$

where  $W_k^{sp} \in \mathbb{R}^{n \times n}$ ,  $b_k^{sp} \in \mathbb{R}^n$  and  $v_k^{sp} \in \mathbb{R}^n$  are the parameters to be learned, and  $n$  is the dynamic feature size of any location. After the context vector  $c_{k,t}$  is computed, it is fed to the recurrent unit.

**Recurrent unit** is used to capture the temporal relationships among the reweighted global dynamic features for the community  $k$  as follows:

$$h_{k,t} = f_k(h_{k,t-1}, c_{k,t}) \quad (3)$$

where  $f_k(\cdot)$  is LSTM [11] for community  $k$ , and  $h_{k,t}$  is the  $t$ -th hidden state of  $k$ -th community. We use LSTM in our model (in each community block) since it addresses the vanishing and exploding gradient problems of basic RNNs.

**Temporal Attention** is applied on top of LSTMs to differentiate the contribution of latent representations of global dynamic features at each time point and for each community. To make the output specific to the target location, we incorporate a *query* scheme based on a time-dependent community membership (i.e., contribution of each location to the community) where the membership is further weighted based on the location's spatial proximity to target location (with nearby locations getting larger weights). Specifically, let  $\beta_k^{(i)}$  denotes the attention weight over the hidden state ( $h_{k,i}$ ) of community  $k$  at time  $i$ . The context vector  $\nu_k$ , which is aggregate contribution from community  $k$ , can be learned through the proximity-based weighting scheme, as follows:

$$q_k^{(t)} = x_d^{prox} \odot \alpha_{k,t}, \quad (4)$$

$$\beta_k^{(i)} = \frac{\exp(q_k^{(i)})}{\sum_{t=1}^w \exp(q_k^{(t)})}, \quad \nu_k = \sum_{t=1}^w \beta_k^{(t)} h_{k,t}, \quad (5)$$

where  $x_d^{prox} \in \mathbb{R}^L$  is a vector encoding the proximity of the target location  $d$  to all locations. Here, the proximity of two locations is calculated based on the inverse of geographic distance (*haversine*):

$$prox(l_1, l_2) = \frac{1}{\sqrt{1 + dist(l_1, l_2)}}. \quad (6)$$

**Community Attention** aims to produce a global contribution with respect to the target location  $d$  by combining different representation subspaces for each of the communities  $\{\nu_1, \nu_2, \dots, \nu_K\}$ . A soft-attention approach is then employed to combine the contributions from all  $K$  communities. Here, to make the prediction specific to the target location, we incorporate a *query* scheme, which takes each community vector  $\{\nu_k\}$  as a *key* and the embedding of the target location as a *query*, as follows:

$$u_k = r^T \tanh(V\nu_k + emb_d), \quad (7)$$

$$\gamma^{(i)} = \frac{\exp(u_i)}{\sum_{k=1}^K \exp(u_k)}, \quad \nu = \sum_{k=1}^K \gamma^{(k)} \nu_k, \quad (8)$$

where  $V \in \mathbb{R}^{m \times m}$ , and  $r \in \mathbb{R}^m$  are the parameters to be learned, where  $m$  is the number of hidden units in LSTMs in the community blocks, and  $\nu$  is the output of the global component.

**Local Component.** This component is designed to model the contribution of the local dynamic features for any target location  $d$  (Fig. 1a). It basically includes a recurrent unit and a temporal attention that focuses on the most informative time instants. The dynamic features of target location are fed to the recurrent unit to model local dynamics.

$$s_t = g(s_{t-1}, X_{t,d}^{dyn}), \quad (9)$$

where  $g(\cdot)$  is LSTM, as in the global component, and  $s_t$  is the  $t$ -th hidden state of LSTM. Then, we also employ a temporal attention on top of the LSTM in this component, which can select the most informative hidden states (time instants) with respect to the dynamic features of target location  $d$ . We only provide the calculation of output vector of the local component to be succinct as:  $\xi_d = \sum_{t=1}^w \delta^{(t)} s_t$  where  $\delta^{(t)}$  is the attention weight for the hidden state at time  $t$ , and  $\xi_d$  is the output of the local component.

**Static Component.** This component models the static information specific to the target location (Fig. 1b). The input incorporates the static features,  $X_d^{stat}$ , and a one-hot encoding vector  $x_d^{id} \in \mathbb{R}^L$  that represents the target location. We apply a fully connected layer (FC) to separately learn a latent representation for each of the two types of information. In particular, the one-hot location vector will be converted into an embedding  $emb_d$  and will be utilized in the aforementioned query component (see Eq. (7)).  $\Psi_d$  is the output of this component, which is concatenation of learned embeddings and latent representation of static features.

**Objective Function.** The objective function consists of three terms: prediction loss, orthogonality loss and Group Lasso (GL) regularization as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{predict} + \lambda \mathcal{L}_{ortho} + \eta \mathcal{L}_{GL}, \quad (10)$$

where  $\lambda$  and  $\eta$  are the tuning parameters for the penalty terms, and  $\mathcal{L}_{predict} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2$ , is the mean squared error (MSE),  $\hat{y}_i$  and  $y_i$  are the predicted and actual number of opioid overdose incidents for sample  $i$ , respectively. A penalty term,  $\mathcal{L}_{ortho}$  is added to avoid learning redundant memberships across communities, i.e., multiple communities may consist of a similar group of locations. To encourage community memberships to be distinguishable as much as possible, we incorporate this orthogonality loss term into the objective function. Let  $\bar{\alpha}_k$  be the community membership vector denoting how each location contributes to the community  $k$ , averaging over time, and  $\Delta = [\bar{\alpha}_1, \bar{\alpha}_2, \dots, \bar{\alpha}_K] \in \mathbb{R}^{K \times L}$  is a matrix consisting of such membership vectors for all communities, the orthogonality loss is given by:

$$\mathcal{L}_{ortho} = \|\Delta \cdot \Delta^T - I\|_F^2, \quad (11)$$

where  $I \in \mathbb{R}^{K \times K}$  is the identity matrix. This loss term encourages different communities to have non-identical locations as members as much as possible, which helps reduce the redundancy across communities. Lastly, we incorporate GL regularization into objective function, which imposes sparsity on a group level [22], and which has been found effective in several domains ([27, 18]) to select informative features. Our main motivation to employ GL is to select community-level and local-level informative features. It enables us to interpret and differentiate which features are important for opioid overdose incidents. It is defined as:

$$\mathcal{L}_{GL} = \sum_{k=1}^K \|Z_k^{glob}\|_{2,1} + \|Z^{local}\|_{2,1} + \|Z^{stat}\|_{2,1}, \quad (12)$$

$$\|Z\|_{2,1} = \sum_{g \in G} \sqrt{|g|} \|g\|_2, \quad (13)$$

where  $Z_k^{glob}$  denotes input weight matrix in the  $k^{th}$  community block in the global component.  $Z^{local}$  and  $Z^{stat}$  represent input weight matrices in the local and the static components, respectively.  $g$  is vector of outgoing connections (weights) from an input neuron,  $G$  denotes a set of input neurons, and  $|g|$  indicates the dimension of  $g$ .

### 3.3 Features

We employ two types of features namely static features and dynamic features.

**Static features** include economical status, housing status, educational level of neighborhoods and demographics such as population, gender diversity index and race diversity index, which are obtained from census data. The diversity index is calculated using

normalized entropy. Furthermore, we employ median household income, per capita income and poverty (%) as the economical indicators. We utilize percentage of the vacant houses (housing occupancy) and percentage of owner occupied houses (housing tenure) as the housing-related static features. We also consider percentage of high school graduation and below as the educational attainment indicator as another static feature. As a result, we obtain a total of nine static features. Note that we apply z-score normalization for median household income and per capita income, and log-transformation for population while preparing the feature vectors.

**Dynamic features** are to capture the crime dynamics of the locations that may be predictive for opioid overdose. We extract the dynamic features from public safety data portals of the cities. Each crime incident is identified by a unique crime incident number and has a certain type which shows a hierarchical structure. The crime data gathered from different cities may have very different categories. For example, the dataset from the City of Chicago includes much more categories than that from the City of Cincinnati. Here, we only consider the highest level, “primary crime types” and eliminate rare categories. The full list of crime categories used in this work can be found in Fig. 4. In addition to these features, we also utilize the number of total crimes and the number of total opioid overdose incidents as additional dynamic features. For each neighborhood and each time unit, the feature vector contains the total number of crimes, the total number of incidents for each primary crime type and the number of opioid overdose. We apply z-score normalization to all dynamic features.

## 4 Experiments

### 4.1 Datasets

We apply our method to forecast opioid overdose on two cities, namely City of Chicago and City of Cincinnati. The neighborhood boundaries officially recognized by the City of Cincinnati and the City of Chicago are called “Statistical Neighborhood Approximations (SNAs)” and “community areas”, respectively. Hereafter, we use “neighborhoods” to refer to both. There are 77 neighborhoods in Chicago whereas Cincinnati consists of 50 neighborhoods. While we select 47 neighborhoods from Chicago (where  $\sim 80\%$  of opioid overdose deaths occur), we use all neighborhoods of Cincinnati in our experiments. Table 1 shows descriptive information about both datasets. For each city, we collect three types of data related to crime, opioid overdose and census as follows:

**Crime data:** We collect crime incident information (geo-location, time and primary type of the crimes) from the open data portals of the cities. We use Public Safety Crime dataset<sup>\*\*\*</sup> and Police Data Initiative (PDI) Crime Incidents dataset<sup>†</sup> to extract such information for City of Chicago and City of Cincinnati, respectively. We extract 14 crime-related dynamic features for Chicago, and 9 dynamic features for Cincinnati.

**Opioid overdose data:** We collect different types of opioid overdose data for each city since there is no systematic monitoring of drug abuse at either a regional or state level in the U.S. For Chicago, we collect opioid overdose death records (geo-location and

<sup>\*\*\*</sup> <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

<sup>†</sup> <https://data.cincinnati-oh.gov/Safer-Streets/PDI-Police-Data-Initiative-Crime-Incidents/k59e-2pvf>



Table 1: Descriptive information about our datasets.

	#NBHDs	#Dynamic Features	#Static Features	#Crimes	#Opioid ODs	Incident Type	Time Interval	Time Span
Chicago	47	15	9	573207	1468	deaths	1 week	08/03/15 08/26/18
Cincinnati	50	10	9	75779	5401	EMS calls	1 week	08/01/15 06/01/18

time) from Opioid Mapping Initiative Open Datasets<sup>‡</sup>. On the other hand, we utilize the EMS response data<sup>§</sup> for heroin overdoses in Cincinnati.

**Census data:** We use the 2010 Census data to extract the features related to demographics, economical status, housing status and educational status of the neighborhoods.

#### 4.2 Baselines

We compare our model with a number of baselines as follows:

- **HA:** Historical average.
- **ARIMA:** is a well-known method for predicting future values for time series.
- **VAR:** captures the linear inter-dependencies among multiple time series and forecasts future values.
- **SVR:** is Support Vector Regression. We use its two variants,  $SVR_{ind}$  (trained separate models for each location) and  $SVR_{all}$  (trained a single model for all locations).
- **LSTM:** We train an LSTM network in which dynamic features are fed to LSTM, then the latent representations are concatenated with static features for prediction.
- **DA-RNN** [20]: is a dual-staged attention-based RNN model for spatio-temporal time series prediction.
- **GeoMAN** [17]: is a multi-level attention-based RNN model for spatio-temporal prediction, which shows state-of-the-art performance in the air quality prediction task.
- **ActAttn** [5]: is a hierarchical spatio-temporal predictive framework for social movements. We replace the final classification layer with regression layer to configure it to regression task to use it as another baseline.

Furthermore, to evaluate the effectiveness of individual components of our model, we also include its several variants for the comparison as follows:

- **CASTNet-noGL:** GL regularization is not incorporated into the loss function.
- **CASTNet-noOrtho:** Orthogonality penalty is not applied so that differentiation of the communities is not encouraged.
- **CASTNet-noSA:** The spatial attentions are removed from the community blocks. Instead, the feature vectors of all locations are concatenated.
- **CASTNet-noTA:** The temporal attentions in both local and global components are removed from the architecture.
- **CASTNet-noCA:** The community attention is removed from the architecture. Instead, the context vectors of the communities are concatenated.
- **CASTNet-noSC:** The static features are excluded from the architecture, yet the location-ID is still embedded.

**Settings:** We used ‘week’ as time unit and ‘neighborhood’ as location unit. We divided datasets into training, validation and test sets with ratio of 75%, 10% and 15%, respec-

<sup>‡</sup> <https://opioidmappinginitiative-opioidepidemic.opendata.arcgis.com/>

<sup>§</sup> <https://insights.cincinnati-oh.gov/stories/s/Heroin/dm3s-ep3u/>

Table 2: Performance Results.

	Chicago		Cincinnati	
	MAE	RMSE	MAE	RMSE
HA	0.2329	0.3385	0.5728	0.8727
ARIMA	0.2272	0.3396	0.5717	0.8952
VAR	0.2242	0.3386	0.5606	0.8712
SVR <sub>ind</sub>	0.2112	0.3321	0.5153	0.8609
SVR <sub>all</sub>	0.1984	0.3063	0.4886	0.8602
LSTM	0.2024	0.3134	0.5235	0.8267
DA-RNN [20]	0.1726	0.3051	0.4817	0.8225
GeoMAN [17]	0.1679	0.2829	0.5034	0.8453
ActAttn [5]	0.1693	0.2937	0.4827	0.8326
CASTNet-noGL	0.1662	0.3129	0.4703	0.8311
CASTNet-noOrtho	0.1649	0.2948	0.4716	0.8109
CASTNet-noSA	0.1608	0.2893	0.4579	0.8152
CASTNet-noTA	0.1641	0.2876	0.4700	0.8141
CASTNet-noCA	0.1631	0.3069	0.4730	0.8225
CASTNet-noSC	0.1693	0.2980	0.4692	0.8291
<b>CASTNet</b>	<b>0.1391</b>	<b>0.2679</b>	<b>0.4516</b>	<b>0.8032</b>

tively. We set  $\tau = 1$  to make short-term predictions. For RNN-based methods, hidden unit size of LSTMs was selected from  $\{8, 16, 32, 64\}$ . The networks were trained using Adam optimizer with a learning rate of 0.001. For each LSTM layer, dropout of 0.1 was applied to prevent overfitting. In our models, the regularization factors  $\lambda$  and  $\eta$  were optimized from the small sets  $\{0.001, 0.005, \dots, 0.05\}$  and  $\{0.001, 0.0015, \dots, 0.01\}$ , respectively using grid search. For *ARIMA* and *VAR*, the orders of the autoregressive and moving average components were optimized for the time lags between 1 and 11. For RNN-based methods, we performed experiments with different window sizes  $w \in \{5, 10, 15, 20\}$ , and shared the results for  $w = 10$  (the best setting for all models). Our code and data are available at <https://github.com/picsoLab/castnet>.

## 5 Results

### 5.1 Performance Comparison

Table 2 shows that CASTNet achieves the best performance in terms of both mean absolute error (MAE) and root mean squared error (RMSE) on both datasets. Our model shows 17.2% and 5.3% improvement in terms of MAE and RMSE, respectively, on Chicago dataset compared to state-of-the-art approach GeoMAN. Similarly, CASTNet enhances the performance 6.3% and 2.4% on Cincinnati dataset in terms of MAE and RMSE, respectively, compared to DA-RNN which shows best performance among the other baselines. Furthermore, we observe that mostly spatio-temporal RNN-based models outperform other baselines, which indicates they better learn the complex spatio-temporal relationships between crime and opioid overdose dynamics.

We further evaluate the effectiveness of each individual component of CASTNet with an ablation study. As described in Section 4.2, each variant is different from the proposed CASTNet by removing one tested component (with others kept identical as much as possible). Table 2 shows that the removal of GL in the model results in a significantly lower performance compared to the others. In addition, CASTNet-noGL can no longer be able to select informative features. Similarly, excluding orthogonality term (CASTNet-noOrtho) loses the ability to learn distinguishable communities or represen-

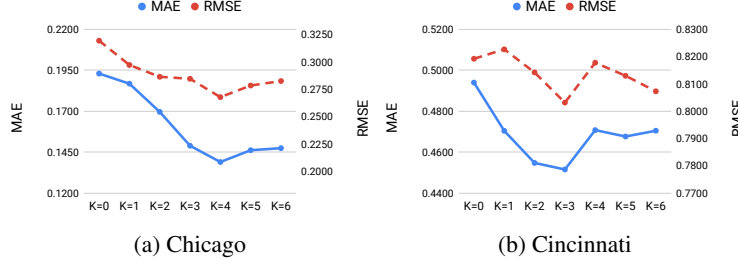


Fig. 2: MAE and RMSE results w.r.t change in the number of communities.

tation subspaces and reduces the performances as well. Moreover, comparing CASTNet with CASTNet-noCA shows that employing community attention has a great impact on the performance, which indicates that learning pairwise activity relationships between a particular event location and communities is crucial. Location-specific static features are also informative since their exclusion (CASTNet-noSC) degrades the performance in both cases. The individual component that provides the least performance gain is spatial attention for both cases. However, its removal (CASTNet-noSA) results in loss of interpretability capability of community memberships. These results reflect that each individual component has important contribution to forecasting performance.

Moreover, we evaluate the performance of the CASTNet with respect to the change in number of communities  $K$ . We conduct experiments with different values of  $K$  selected from  $\{0, 1, \dots, 6\}$  and the results are given in Fig. 2. Note that the model does not consider global contribution when  $K = 0$ . Also, when  $K = 1$ , the model yields a single universal representation of global activities which is irrespective of the event locations. The best performances are obtained when  $K = 4$  for Chicago and  $K = 3$  for Cincinnati datasets. We observe that while  $K$  increases until the optimum value, the performance increases, and some communities are decomposed to form new communities. However, as long as  $K$  continues to increase after its optimum value, the performance starts to decrease slightly or remains stable, and the semantic subspaces of some communities become similar. With this experiment, we indicate that learning different representations of global activities significantly improves the forecasting performance.

## 5.2 Analysis of Community Memberships and Community Contributions

We analyze the community memberships of the neighborhoods and community contributions on forecasting future opioid overdose by answering the following questions. **How do locations contribute to communities?** CASTNet learns different representation subspaces (communities) of global dynamics unlike the previous work [17, 5], and each community consists of a group of different members due to orthogonality penalty. We represent the learned communities and their memberships (i.e., the spatial attention weights  $\alpha$  in Eq. (2), averaged over time for ease of interpretation) on the left side of Fig. 3a and 3b for Chicago and Cincinnati, respectively, where the line thickness represents the degree at which a location contributes to the corresponding communities. Note that neighborhoods on the left side of Fig. 3a and Fig. 3b are ordered by the number of crimes. As shown in Fig. 3, most locations have dedicated to one community. For Chicago model (Fig. 3a), *Austin* (25), which has the highest number of crime incidents and opioid overdose deaths, formed a separate community  $C_4$  by itself. While *North Lawndale* (29) and *Humboldt Park* (23) together formed the community  $C_1$ , *West Garfield Park* (26), *East Garfield Park* (27) and *North Lawndale* (29) formed an another

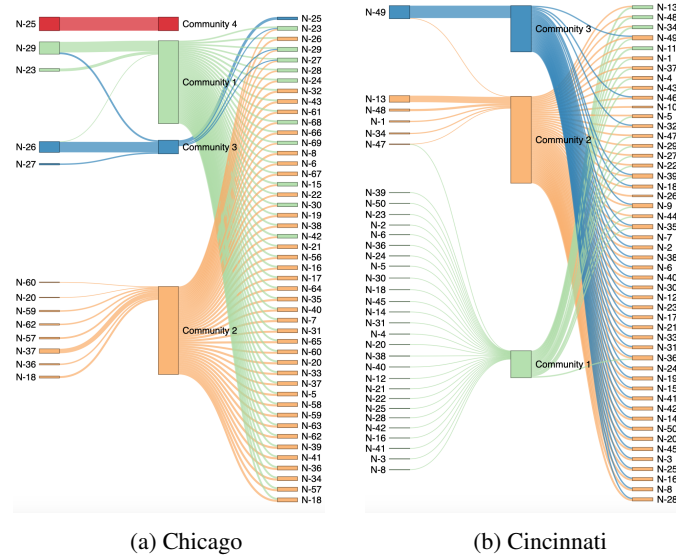


Fig. 3: **Community memberships and community contributions on forecasting.** For each community, left side represents community memberships (how each location contributes to the community), and right side represents the average community contribution (how the community contribute to predicting a target location). Edge thickness indicates the weight of community membership (left side) and community contribution (right side). Node size denotes overall community membership of a location (left side) and overall community contribution to forecasting overdose in the target neighborhood (right side). Edge color shows the input and output of a specific community. Node color of a neighborhood indicates the community for which the corresponding neighborhood has the highest membership (left side). Node color of a neighborhood denotes the community from which the neighborhood takes the largest contribution (right side). Edges whose weights are above a certain threshold are shown.

community  $C_3$ . Note that neighborhoods of  $C_1$  and  $C_3$  have the highest opioid overdose death rate after *Austin* (25). On the other hand, the community  $C_2$  is formed by the neighborhoods having low crime and overdose death rates including *Fuller Park* (37), *McKinley Park* (59) and *West Elsdon* (62). Furthermore, for Cincinnati model (Fig. 3b), *Westwood* (49), where the highest number of crimes were committed, formed a separate community  $C_3$  by itself. It shows a similar behavior to the Chicago case. *East Price Hill* (13), *West Price Hill* (48), *Avondale* (1) and *Over-The-Rhine* (34) formed the community  $C_2$  where these neighborhoods have the highest crime rate after *Westwood* (49) and the highest opioid overdose rate. On the other hand, the community  $C_1$  is formed by rest of the neighborhoods (with low and moderate crime rates) and their memberships of that community are almost equal.

**How do the communities contribute to forecasting?** CASTNet is capable of modeling the pairwise activity relationships between a particular event location and the communities. It allows the target location to attend the communities to select location-specific global contributions. We analyze how these communities contribute to forecast-

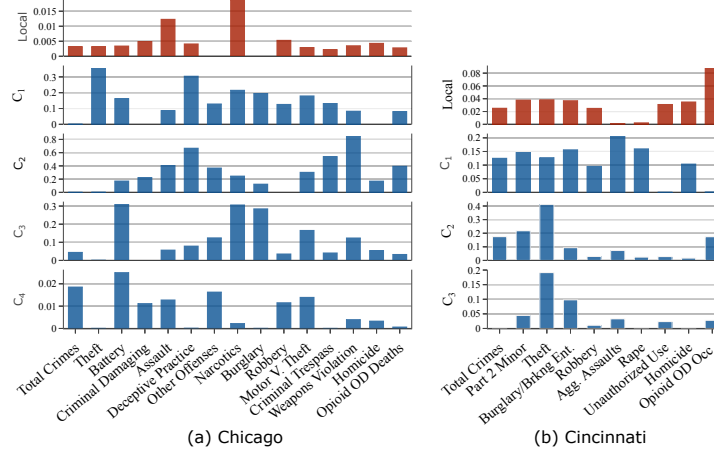


Fig. 4: **Importance of dynamic features.** Mean absolute values of input weights of local and global components.

ing by visualizing the community attention weights (i.e.,  $\gamma$  in Eq. (8) averaged over test samples for each neighborhood) in Fig. 3a and Fig. 3b for Chicago and Cincinnati, respectively. While the left side of the figures represents the community memberships, the right side indicates the average community contributions for each neighborhood. Note that neighborhoods on the right side of Fig. 3a and Fig. 3b are ordered by the number of opioid overdoses. For Chicago case,  $C_1$  and  $C_2$  have more contributions than the other communities on forecasting overdose. While  $C_2$  contributes more to neighborhoods with low or moderate opioid overdose death rate,  $C_1$  contributes more to the neighborhoods where the death rate is higher.  $C_3$  also contributes more to the neighborhoods with the highest death rate (e.g. *Austin* (25), *Humboldt Park* (23)). This means that any particular neighborhood attends more to the community, which is formed by the similar neighborhoods. On the other hand,  $C_4$  does not significantly contribute to any neighborhood although it is formed by a crime hot-spot (*Austin* (25)). For Cincinnati case,  $C_2$  is a very dominant community, which makes the largest global contribution to most of the neighborhoods. The neighborhoods that formed  $C_2$  and  $C_3$  (e.g. *East Price Hill* (13), *West Price Hill* (48), *Westwood* (49)) are very predictive, and the change in their dynamics have greater impact on forecasting future overdoses in the target neighborhoods. On the other hand,  $C_1$  has larger contribution to neighborhoods where the overdose rate is the highest. This indicates that the crimes committed in the members of  $C_1$  are also informative for forecasting future overdoses in opioid hot-spots.

### 5.3 Feature Analysis

We investigate the importance of **dynamic features** by analyzing the mean absolute input weights of local and global components as shown in Fig. 4. For Chicago case, GL selects *Narcotics* and *Assault* as the most important features for future opioid overdose deaths in the same location. Moreover, *Theft*, *Deceptive Practice*, *Narcotics*, *Burglary* and *Motor V. Theft* are the predictive features from  $C_1$  while *Weapons Violation*, *Deceptive Practice* (e.g. *Fraud*) and *Criminal Trespass* are significant from  $C_2$ . Recall that,  $C_1$  and  $C_2$  are the most contributing communities to forecasting (see Fig. 3a). This

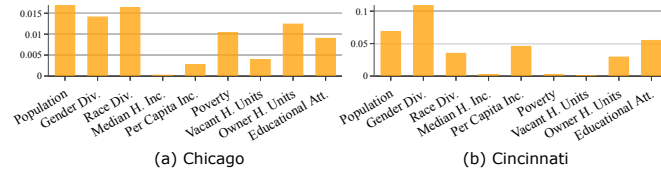


Fig. 5: **Importance of static features.** Mean absolute values of input weights of FC layer in static component.

shows that property crimes (e.g. *Theft*, *Burglary*, *Deceptive Practice*) are more significant predictors than the violent crimes for Chicago. Such crimes previously committed in the members of  $C_1$  and  $C_2$  may be a significant indicator of future opioid overdose deaths in Chicago. On the other hand, *Battery*, *Narcotics*, *Burglary*, and *Motor V. Theft* are predictive features from  $C_3$  while *Battery*, *Total Crimes* and *Other Offenses* (e.g. *offenses against family*) are significant from  $C_4$ . However,  $C_3$  has larger contribution than other communities for only *Austin* (25).  $C_4$  does not provide a significant contribution to any neighborhood as much as the other communities. For Cincinnati case, *Opioid Overdose Occ.* is the most predictive feature for forecasting future opioid overdose in the same location, which means the local component behaves as an autoregressive module unlike the Chicago case. Furthermore, both violent crimes including *Agg. Assaults*, *Rape*, *Homicide*, *Part 2 Minor* (e.g. *Menacing*) and property crimes including *Burglary/Breaking Ent.*, *Theft*, *Part 2 Minor* (e.g. *Fraud*) are significant features from  $C_1$ . On the other hand, *Theft* and *Part 2 Minor* from  $C_2$ , and *Theft* and *Burglary* from  $C_3$  are predictive features for future opioid overdose in the target locations. Recall that  $C_2$  and  $C_3$  have more salient contribution on most of the neighborhoods, which implies that commitment of previous property crimes (especially *Theft*) in the members of those communities may be one of the potential indicators of future opioid overdose in the other neighborhoods. Note that our findings are also consistent with the literature that highlighted the connection between crime and drug use, and suggested the property crimes such as theft, burglary might be committed to raise funds to purchase drugs [1].

We explore the importance of **static features** by analyzing mean absolute input weights of FC in static component (see Fig. 5). For Chicago case, demographic features (*Population*, *Gender Div.* and *Race Div.*) are significant. We observe that *Owner Occupied H. units*, *Poverty* and *Educational Att.* are also informative. For Cincinnati case, *Gender Div.* and *Population* are important features for forecasting as well as the *Educational Att.* and *Per Capita Income*. Based on these results, the neighborhoods with higher population, and lower or moderate gender diversity may require additional resources to prevent opioid overdose in both cities. Also, economic status is important for neighborhoods of both cities, which is consistent with the previous work that suggested communities with a higher concentration of economic stressors (e.g. low income, poverty) may be vulnerable to abuse of opioids as a way to manage chronic stress and mood disorders [15]. Although there exist three economic status indicators, GL selects only one, *Poverty* for Chicago and *Per Capita Income* for Cincinnati.

## 6 Discussion and Future Work

In this work, we presented a community-attentive spatio-temporal predictive model to forecast opioid overdose from crime dynamics. We developed a novel deep learn-

ing architecture based on multi-head attentional networks that learns different representation subspaces of features (communities) and allows the target locations to select location-specific community contributions for forecasting local incidents. At the same time, our proposed model allows for interpreting predictive features in both local-level and community-level, as well as the community memberships and the community contributions to forecasting local incidents. We demonstrated the strength of our method through the extensive experiments. Our method achieved superior forecasting performance on two real-world opioid overdose datasets compared to the baseline methods.

The experiment results suggest different spatio-temporal crime-overdose potential links. The overdose deaths at a target neighborhood in Chicago appear to be better predicted by crime incidents at neighborhoods that share the same community with the target neighborhood. Also, change in crime incidences in those neighborhoods with low crime rates is an important indicator of future overdose deaths in most of the other neighborhoods. On the other hand, in Cincinnati, the crime incidents occurred in communities comprising those crime hot-spots seem to well predict the overdose events in most of the neighborhoods. Furthermore, the predictive local activities are different in two cases. While the local crime incidents, in particularly *Narcotics* and *Assault*, are predictive for local overdose deaths in Chicago, previous overdose occurrences are informative for future overdose incidents in Cincinnati. On the other hand, the global contributions to forecasting local overdose incidents show similar patterns in both cities. Change in property crimes, in particular *Theft*, *Deceptive Practice*, *Burglary* and *Weapons Violation* (crime against to society) in Chicago, *Theft* and *Burglary* in Cincinnati, can be significant indicators for future local overdose incidents as well as certain type of violent crimes (*Battery* for Chicago and *Agg. Assault* for Cincinnati). Last but not the least, demographic characteristics, economic status and educational attainment of the neighborhoods in both cities may help forecasting future local incidents. Our findings support the hypothesis that criminal activities and opioid overdose incidents may reveal spatio-temporal lag effects, and are consistent with the literature. As future work, we plan to investigate the link between opioid use (or overdose) and other social phenomena using our method. We also plan to extend our model to consider multi-resolution spatio-temporal dynamics for prediction.

**Acknowledgement.** This work is part of the research associated with NSF #1637067 and #1739413. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the funding sources.

## References

1. Bennett, T., Holloway, K., Farrington, D.: The statistical association between drug misuse and crime: A meta-analysis. *Aggression and Violent Behavior* **13**(2), 107–118 (2008)
2. Burke, D.S.: Forecasting the opioid epidemic. *Science* (354), 529 (2016)
3. Centers for Disease Control and Prevention: Web-based Injury Statistics Query and Reporting System (WISQARS). <http://www.cdc.gov/injury/wisqars/fatal.html>. Last accessed: 2018-05-01
4. Chalapathy, R., Toth, E., Chawla, S.: Group anomaly detection using deep generative models. In: ECML-PKDD. pp. 173–189 (2018)
5. Ertugrul, A.M., Lin, Y.R., Chung, W.T., Yan, M., Li, A.: Activism via attention: interpretable spatiotemporal learning to forecast protest activities. *EPJ Data Science* **8**(1), 5 (2019)

6. Ertugrul, A.M., Lin, Y.R., Mair, C., Taskaya Temizel, T.: Forecasting heroin overdose occurrences from crime incidents. In: SBP-BRIMS (2018)
7. Ghaderi, A., Sanandaji, B.M., Ghaderi, F.: Deep forecast: Deep learning-based spatio-temporal forecasting. arXiv preprint arXiv:1707.08110 (2017)
8. Glanz, J.M., Narwaney, K.J., Mueller, S.R., Gardner, E.M., Calcaterra, S.L., Xu, S., Breslin, K., Binswanger, I.A.: Prediction model for two-year risk of opioid overdose among patients prescribed chronic opioid therapy. *J. of general internal medicine* **33**(10), 1646–1653 (2018)
9. Gruenewald, P.J.: Geospatial analyses of alcohol and drug problems: empirical needs and theoretical foundations. *GeoJournal* **78**(3), 443–450 (2013)
10. Hammersley, R., Forsyth, A., Morrison, V., Davies, J.B.: The relationship between crime and opioid use. *Addiction* **84**(9), 1029–1043 (1989)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
12. Huang, C., Zhang, J., Zheng, Y., Chawla, N.V.: Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In: ACM CIKM. pp. 1423–1432 (2018)
13. Jalal, H., Buchanich, J.M., Roberts, M.S., Balmert, L.C., Zhang, K., Burke, D.S.: Changing dynamics of the drug overdose epidemic in the united states from 1979 through 2016. *Science* **361**(6408), eaau1184 (2018)
14. Kennedy-Hendricks, A., Richey, M., McGinty, E.E., Stuart, E.A., Barry, C.L., Webster, D.W.: Opioid overdose deaths and florida’s crackdown on pill mills. *American journal of public health* **106**(2), 291–297 (2016)
15. King, N.B., Fraser, V., Boikos, C., Richardson, R., Harper, S.: Determinants of increased opioid-related mortality in the united states and canada, 1990–2013: a systematic review. *American journal of public health* **104**(8), e32–e42 (2014)
16. Kolodny, A., Courtwright, D.T., Hwang, C.S., Kreiner, P., Eadie, J.L., Clark, T.W., Alexander, G.C.: The prescription opioid and heroin crisis: a public health approach to an epidemic of addiction. *Annual review of public health* **36**, 559–574 (2015)
17. Liang, Y., Ke, S., Zhang, J., Yi, X., Zheng, Y.: Geoman: Multi-level attention networks for geo-sensory time series prediction. In: IJCAI. pp. 3428–3434 (2018)
18. Ochiai, T., Matsuda, S., Watanabe, H., Katagiri, S.: Automatic node selection for deep neural networks using group lasso regularization. In: ICASSP (2017)
19. Pierce, M., Hayhurst, K., Bird, S.M., Hickman, M., Seddon, T., Dunn, G., Millar, T.: Quantifying crime associated with drug use among a large cohort of sanctioned offenders in england and wales. *Drug & Alcohol Dependence* **155**, 52–59 (2015)
20. Qin, Y., Song, D., Cheng, H., Cheng, W., Jiang, G., Cottrell, G.W.: A dual-stage attention-based recurrent neural network for time series prediction. In: AAAI. pp. 2627–2633 (2017)
21. Rudd, R.A., Aleshire, N., Zibbell, J.E., Matthew Gladden, R.: Increases in drug and opioid overdose deaths—united states, 2000–2014. *American Journal of Transplantation* **16**(4), 1323–1327 (2016)
22. Scardapane, S., Comminiello, D., Hussain, A., Uncini, A.: Group sparse regularization for deep neural networks. *Neurocomputing* **241**, 81–89 (2017)
23. Seddon, T.: Drugs, crime and social exclusion: social context and social theory in british drugs–crime research. *British Journal of Criminology* **46**(4), 680–703 (2005)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
25. Warner, M., Chen, L.H., Makuc, D.M., Anderson, R.N., Miniño, A.M.: Drug poisoning deaths in the united states, 1980–2008. *NCHS data brief* (81), 1–8 (2011)
26. Zhao, L., Wang, J., Guo, X.: Distant-supervision of heterogeneous multitask learning for social event forecasting with multilingual indicators. In: AAAI (2018)
27. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: AAAI (2016)