Adversarial Adaptation of Scene Graph Models for Understanding Civic Issues

Conference Paper · May 2019

DOI: 10.1145/3308558.3313681

4 authors, including:

Shanu Kumar
Indian Institute of Technology Kanpur
5 PUBLICATIONS **136** CITATIONS
SEE PROFILE

Anjali Singh
IBM
6 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Mohit Jain
University of Toronto
30 PUBLICATIONS **1,053** CITATIONS
SEE PROFILE

Adversarial Adaptation of Scene Graph Models for Understanding Civic Issues

Shanu Kumar
Indian Institute of Technology Kanpur, India
sshanu@iitk.ac.in

Shubham Atreja
IBM Research AI, India
satreja1@in.ibm.com

Anjali Singh
IBM Research AI, India
ansingh8@in.ibm.com

Mohit Jain
IBM Research AI, India
mohitjain@in.ibm.com

ABSTRACT

Citizen engagement and technology usage are two emerging trends driven by smart city initiatives. Typically, citizens report issues, such as broken roads, garbage dumps, *etc.* through web portals and mobile apps, in order for the government authorities to take appropriate actions. Several mediums – text, image, audio, video – are used to report these issues. Through a user study with 13 citizens and 3 authorities, we found that image is the most preferred medium to report civic issues. However, analyzing civic issue related images is challenging for the authorities as it requires manual effort. In this work, given an image, we propose to generate a *Civic Issue Graph* consisting of a set of objects and the semantic relations between them, which are representative of the underlying civic issue. We also release two multi-modal (text and images) datasets, that can help in further analysis of civic issues from images. We present an approach for adversarial adaptation of existing scene graph models that enables the use of scene graphs for new applications in the absence of any labelled training data. We conduct several experiments to analyze the efficacy of our approach, and using human evaluation, we establish the appropriateness of our model at representing different civic issues.

KEYWORDS

Civic Engagement, Scene Graph Generation, Adversarial Adaptation, Smart Cities, Intelligent Systems on Web

ACM Reference Format:
Shanu Kumar, Shubham Atreja, Anjali Singh, and Mohit Jain. 2019. Adversarial Adaptation of Scene Graph Models for Understanding Civic Issues. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313681

1 INTRODUCTION

In recent years, there has been a significant increase in smart city initiatives [8, 26, 27]. As a result, government authorities are emphasizing the use of technology and increased citizen participation for better maintenance of urban areas. Various web platforms – SeeClickFix [23], FixMyStreet [1], ichangemycity [13] – have been introduced across the world, which enable the citizens to report civic issues such as poor road condition, garbage dumps, *etc.*, and track the status of their complaints. This has resulted in exponential increase in the number of civic issues being reported [2]. Even social media sites (Twitter, Facebook) have been increasingly utilized to report civic issues. These issues are reported online through various mediums – textual descriptions, images, videos, or a combination of them. Previous work [9] highlights the importance of different mediums in citizen participation. Yet, no prior work has tried to understand the role of these mediums in reporting of civic issues.

In this work, we first identify the most preferred medium for reporting civic issues, by conducting a user study with 13 citizens and 3 government authorities. Using the 84 civic issues reported by the citizens using our mobile app, and follow-up semi-structured interviews, we found that images are the most usable medium for the citizens. In contrast, authorities found text as the most preferred medium, as images are hard to analyze at scale.

To fill this gap, several works have proposed methods to automatically identify a specific category of civic issues from images, such as garbage dumps [24] and road damage [20]. However, their methods are limited to the specific categories that they address.

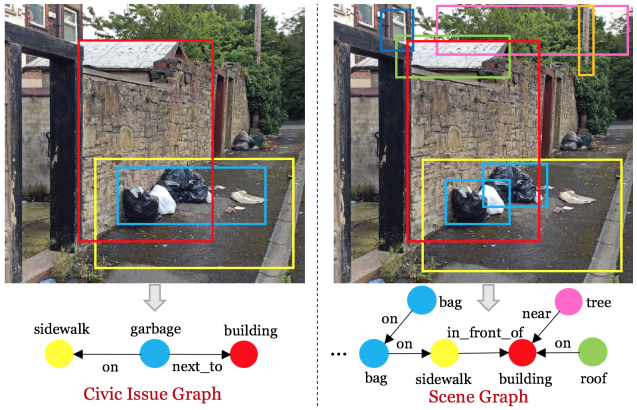

Figure 1: Comparison between Civic Issue Graph and Scene Graph for a given image. The scene graph provides a complete representation of all objects and relationships in the image, while the Civic Issue Graph only consists of those objects and relations which are representative of the civic issue.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313681

Furthermore, existing holistic approaches of analyzing civic issues are limited to text [4]. To this end, we propose an approach to understand various civic issues from input images, independent of the type of issue being reported.

In the field of image understanding, scene graphs [14], are used to get a complete representation of all objects in an image along with the relations between them. However, to understand a civic issue, only certain crucial objects and the relations between them need to be detected, which are representative of the civic issue in the image. Inspired from the task of scene graph generation, we propose to generate *Civic Issue Graphs* that provide complete representations of civic issues in images.

Figure 1 shows a comparison between the two representations. In contrast to the scene graph, the Civic Issue Graph only consists of objects conveying a civic issue, their bounding boxes, and the relation between these objects. We present a formal definition of this representation in Section 5.

Training a scene graph model requires a large amount of data consisting of images with grounded annotations (objects and relations in the images). Due to the lack of sufficient annotated images of civic issues, we use an existing scene graph model in a cross-domain setting, with partially annotated and unpaired data. We utilize a dataset extracted by collating and processing images from civic issue complaint forums, for training our model, and make this dataset publicly available. We present a novel adversarial approach that uses an existing scene graph model for a new task in the absence of any labelled training data. We conduct various experiments to establish the efficacy of our approach using metrics derived from standard scene graph evaluation metrics. Finally, through human evaluation, we demonstrate that civic issues from images can be appropriately represented using our Civic Issue Graph.

To summarize, the major contributions of this paper are: (i) understanding the usability of different mediums for reporting civic issues, (ii) introducing an unsupervised mechanism using adversarial adaptation of existing scene graph models to a new domain, (iii) experimental evaluation indicating significant performance gains for identification of civic issues from user uploaded images, and (iv) releasing two multi-modal (text and image) datasets with information on civic issues, to encourage future work in this domain.

2 RELATED WORK

Civic Issue Detection and Analysis. Social media provides a convenient interface that allows citizens to report civic issues [3, 15]. Several works try to analyze online platforms to automatically mine issues related to civic amenities [22, 25], but the analysis is limited to textual descriptions. Specific to images, [20] and [24] use object detection and image segmentation techniques to identify road damage and garbage dumps respectively, from input images. However, their methods are also limited to the specific category of civic issues that they address. One of the more recent works, 'Citicafe' [4] employs machine learning techniques to understand and analyze different types of civic issue from the user input. However, they do not provide a method for understanding images reporting civic issues, as we do in this paper.

Scene-Graph Generation. Several works [16, 33] propose methods for generating scene graphs from images to represent all objects in an image and the relationships between them. Zellers et al. [33] present the state-of-the-art for generating scene graphs by establishing that several structural patterns exist in scene-graphs (which they call motifs) and showing how object labels are highly predictive of relation labels by analyzing the Visual Genome dataset [17]. However, their approach requires a large set of images for training with grounded annotations for objects and relations. Some works [19, 34] utilize zero shot learning for generating a scene-graph. However, their results show that the learning is restricted to the task of detecting new predicates which were not seen during the training phase. Our approach can be used to generalize existing scene graph models to predict new relations belonging to a different domain, which are absent from the training data.

Domain Adaptation. Domain adaptation is a long studied problem, where approaches range from fine-tuning networks with target data [28] to adversarial domain adaptation methods [31]. Adversarial methods [10, 29, 31] employ a domain classifier to learn mappings from the source domain to target domain, and these mappings are used to generalize the model to the target domain. These methods have shown promising results for image understanding tasks such as captioning [5] and object detection [7]. In this work, we propose to use Adversarial Discriminative Domain Adaptation (ADDA) [31] for adapting scene graph models to our new task.

3 USER STUDY

We conducted a user study to understand the preference of different mediums – text, audio, image and video – to report civic issues, both from citizens and authorities perspective. For this, we developed a custom Android app, with the landing page having four buttons, each corresponding to the four mediums. To report an issue, any of the medium(s) could be used any number of times, *e.g.,* a report can comprise of 1 video, few lines of text, and 2 images.

13 participants (9 male, 4 female, age=28.5±6.1 years) reported civic issues over a period of 7-10 days. All the participants were recruited using word-of-mouth and snowball sampling. All of them were experienced smartphone users, using it for the past 6.2±2.2 years, and well educated (highest education: 1 high school, 3 Bachelors, 6 Masters, 4 PhDs). However, only two of them have previously reported civic issues on online web portals. At the end of the study, a 30-mins semi-structured interview was conducted, to delve deeper into the reasons for (not) using specific medium(s).

Furthermore, we interacted with 3 government authorities (3 male, age = 35-45 years) for 30-mins each, to understand their perspective on the medium of the received complaints. All interviews were audio-recorded, and later transcribed for analysis.

Results: Overall, 84 (6±3.7) civic issues were reported by the 13 participants, mainly in the category of garbage (11/13 participants), potholes (9), blocked sidewalk (6), traffic (5), illegal car parking (3), and stray dogs (3). 81 of these issues consisted of image, text, or their combination, while only 2 had audio and 1 had video. Hence, we only focus on image and text as preferred mediums.

A majority of the participants (10/13) found images the best for reporting civic issues, followed by text (2/13) and video (1/13). Images were preferred primarily as it is quick and easy to click an image, and they convey a lot of information: "*An image is worth 1000 words.*"-P_4, "*its super quick to take pics... even when I pause at a traffic signal, I can take a pic*"-P_{10}. Participants also felt that images are best for conveying the severity of an issue. Interestingly, participants thought that images can "*act as a proof of the problem... images don't lie*"-P_6. Participants felt that people might '*bluff*'/'*exaggerate*' when reporting issues using text. However, participants complained that images can not be used to capture the temporal variations of civic problem, *e.g.*, "*images can't say tell you for how long the garbage has been lying there*"-P_2. For reporting the temporal variations, participants favored text. But participants also found that texting requires more time and effort, compared to clicking images.

When participants were asked to choose the best combination of mediums for reporting civic issues, majority of them (9/14) chose image with text. The combination allows them the freedom to show severity and truthfulness of the issue using image(s), along with adding other details in text.

Following this, we interviewed 3 government authorities, and found about the process of human annotators analyzing the received images to generate captions describing the issue. These captions are then passed on to the relevant authority in writing or via phone calls to take appropriate actions. Also the authorities confirm that a majority of the received complaints comprise of images. However, these images never reach them due to lack adequate technological infrastructure. This confirms that image is the most preferred medium for users, but authorities rely only on textual complaints. To bridge this gap, in this work, we generate text-based descriptions of images that are used for reporting civic issues.

4 DATASET

An extensive dataset of images with annotations for a wide variety of civic issues is currently unavailable. To this end, we mined 485,927 complaints (with 131,020 images) from two civic issue reporting forums – FixMyStreet [1] and ichangemycity [13]. We use them to generate two datasets.

Dataset-1 consists of human-annotated images with the bounding boxes and object labels for 4 object categories – *garbage, manhole, pothole, water logging*. Some of these object categories are not present in any publicly available image datasets. We utilize the annotations from two existing datasets for garbage [24] and potholes [20], and add new images representative of the new object categories along with their annotations, to build Dataset-1, with a total of 1505 images and 2302 bounding boxes.

Dataset-2 consists of examples of Civic Issue Graphs, represented through triples of the form $[object1, predicate, object2]$, specifying the relationship (*predicate*) between a pair of objects (*object*1 and *object*2). We use natural language processing techniques [21, 30] to extract these triples from complaint descriptions. We manually define a set of 19 target object categories which are relevant to the civic domain and map the objects from these triples to our set of target objects based on semantic similarity [12]. We retain only those triples where the predicate defines positional relations (manually determined) and for which both objects are matched with a similarity value greater than 0.4. This dataset consists of 44,353 Civic Issue Graphs, where 8204 are paired with images. There are total 5799 unique relations with 19 object classes and 183 predicate classes. These datasets can be found here[1].

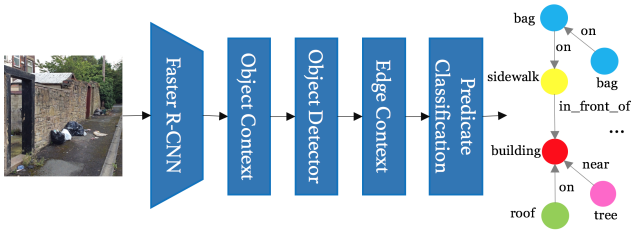

Figure 2: An overview of the MotifNet model

5 CIVIC ISSUE GRAPH GENERATION

We now present our approach for understanding civic issues from input images. We first present the formal definition of Civic Issue Graphs, followed by our detailed approach, consisting of scene graph generation and adversarial domain adaptation.

Formal Definition: A scene graph is a structured representation of objects and the relationships present between them in an image. It consists of triples or relations (used inter-changeably) of the form $[object1, predicate, object2]$ where *predicate* defines the relationship between the two objects and both *object*1 and *object*2 are grounded to their respective bounding box representations in the image. While a scene graph provides a complete representation of the contents of the scene in an image, our proposed Civic Issue Graph (*CG*) only consists of objects conveying a civic issue, their bounding boxes, and the predicate between these objects. We use the following notations to define a *CG*:

- $B = \{b_1, \ldots, b_n\}$: Set of bounding boxes $b_i \in \mathbb{R}^4$;
- $O^1 = \{o^1_1, \ldots, o^1_n\}$: Set of objects essential for defining a civic issue, *e.g.*, '*pothole*', '*garbage*', *etc.*
- $O^2 = \{o^2_1, \ldots, o^2_n\}$: Set of objects that define the context of objects in O^1, *e.g.*, '*street*', '*building*', *etc.*
- $O = O^1 \cup O^2$: Set of all objects that assign a class label $o_i \in O$ to each b_i
- $P = \{p_1, \ldots, p_n\}$: Set of predicates defining geometric or position-based relationships between $o^1_i \in O^1$ and $o^2_i \in O^2$, *e.g.*, '*above*', '*next_to*', '*in*', *etc.*
- $R_{CG} = \{r_1, \ldots, r_n\}$: Set of *CG* relations with nodes $(b_i, o^1_i) \in B \times O^1$, $(b_j, o^2_j) \in B \times O^2$, and predicate label $p_{i \to j} \in P$, *e.g.*, $[garbage, on, street]$, where *'garbage'* $\in O^1$, *'street'* $\in O^2$, and *'on'* $\in P$

5.1 Scene Graph Generation

The MotifNet model, proposed by Zellers et al. [33], is the current state-of-the-art for generating scene graphs and we utilize this model for demonstrating our approach. However, the approach is generic and can be applied to other models with similar architecture.

[1] https://github.com/Sshanu/civic_issue_dataset

MotifNet Model: Fig 2 presents a high-level overview of the MotifNet model. The model consists of an *Object Detector* that provides object labels and their bounding regions in the input image. Following this, the *Object Context* module uses bidirectional LSTMs to generate contextualized representation for each object, and this representation is used by the *Object Decoder* to predict the final object labels. The *Edge Context* module then generates the contextualized representation for each object pair using additional bidirectional LSTMs. Finally, *Predicate Classification* contains a softmax layer to identify the predicate label, using the object-pair representation as input. The reader may refer to the extended version of this work [18] or the original MotifNet paper [33] for a detailed explanation about the individual modules and their mathematical formulations.

5.2 Adversarial Domain Adaptation

Domain adaptation involves using an existing model trained on 'source' domain where labelled data is available, and generalizing it to a 'target' domain, where labelled data is not available. Domain adaptation has been helpful for tasks such as image captioning [6] that require a large corpora of images and their labels, as getting this data for each and every domain is unfeasible. More recently, adversarial methods for domain adaptation [31] have also been proposed, where the training procedure is similar to the training of Generative Adversarial Networks (GANs) [11]. We present an adversarial training approach for a scene graph model, which, to the best of our knowledge, has not been explored before. Domain adaptation for scene graphs is challenging due to the large domain shift in the images as well as the feature space of relations. For instance, the Visual Genome dataset (VG) [17] used for training scene graph models, consists of a mix of indoor and outdoor scenes with more object instances, whereas our dataset of civic issues consists of specific outdoor scenes depicting a civic issue. Moreover, some of the relations observed in the civic issue domain are not even present in the visual genome dataset (e.g., $[garbage, on, street]$). In the following subsections, we provide more details about our cross-domain setting followed by our approach for adversarial domain adaptation.

5.2.1 Cross-Domain Setting. Scene graph models trained on a particular dataset can detect only those relations that are already *seen* by the model, or in other words, present in the training dataset. For our task of generating *CG*, the model needs to detect R_{CG}, *i.e.*, the set of relations contained in *CG*. Note that the set of relations in R_{CG} can be further divided into R_s and R_n, where R_s is the set of relations previously *seen* by the model, *e.g.*: $[tree, over, fence]$ and R_n is the set of relations previously *unseen* by the model *e.g.*: $[garbage, on, street]$. In the absence of any labelled data for R_n, we want to generalize the model trained on R_s, to adapt to R_n as well.

5.2.2 Adversarial Approach. Adversarial approach for domain adaptation consists of two models – a pre-trained generator model and a discriminator model. In our setting, we use the MotifNet model pre-trained on VG dataset as the generator and propose a discriminator model that can distinguish between R_s and R_n. During pre-training, the MotifNet model learns a representation for the object pairs which is used to predict the final set of relations (R_s). Without adversarial training, the model has not learned the representation for any *unseen* pair of objects from the civic domain and will not be able to predict such relations (R_n). Therefore, during adversarial training, the objective of the MotifNet model is to learn a mapping of target object pairs (*unseen*) to the feature space of the source object pairs (*seen*). This objective is supported via the discriminator, which is a binary classifier between the source and target domains. The MotifNet model can be said to have learned a uniform representation of object pairs corresponding to R_s and R_n, if the classifier trained using this representation can no longer distinguish between R_s and R_n. Therefore, we introduce two constrained objectives which seek to – i) find the best discriminator model that can accurately classify R_s and R_n, and ii) "maximally confuse" the discriminator model by learning new mapping for R_n. Once the source and target feature spaces are regularized, the predicate classifier trained on the *seen* object pairs can be directly applied to *unseen* object pairs, thereby eliminating the need for labelled training data.

Fig 3 summarizes our adversarial training procedure. We first pre-train the MotifNet model on the VG dataset using cross-entropy loss and then update it using adversarial training. During adversarial training, the parameters for the MotifNet model and the discriminator are optimised according to a constrained adversarial objective. To optimize the discriminator model, we use the standard classification loss (L_d). In order to optimize the MotifNet model, we use the standard loss function (L_a) with inverted labels (seen $\rightarrow$ unseen, unseen $\rightarrow$ seen) thereby satisfying the adversarial objective. This entire training process is similar to the setting of GANs. We iteratively update the MotifNet model and the Discriminator with a ratio of N_m:N_d with $N_m < N_d$, *i.e.*, the Discriminator is updated more often than the MotifNet model. We now provide a mathematical formulation of our training approach.

Discriminator We define the Discriminator as a binary classifier with *seen* and *unseen* as the two set of classes. For each object pair (o_i, o_j), the Discriminator is provided with two inputs: 1) $\mathbf{g}_{i,j}$: final representation of the object pair generated by the model and 2) $(\mathbf{W}_h \mathbf{d}_i) \circ (\mathbf{W}_t \mathbf{d}_j)$: contextualized representation of the object pair without the visual features. We further experimented with different inputs to the discriminator (details provided in the extended version of this work [18]). The Discriminator consists of 2 fully connected layers, followed by a softmax layer to generate probability $C_d(l|o_i, o_j)$, where $l \in \{seen, unseen\}$. The mathematical formulation of the discriminator for a given object pair (o_i, o_j) is:

$$\mathbf{F}_{i,j} = \mathbf{Dis}([\mathbf{g_{i,j}}; (\mathbf{W_h d_i}) \circ (\mathbf{W_t d_j})]) \tag{1}$$

$$\mathbf{C}_d = softmax(\mathbf{W}_d \mathbf{F}_{i,j} + \mathbf{b}_d) \tag{2}$$

Training Discriminator Let OP_{cv} be the set of all object pairs identified by the model for an image, I_{CV} belonging to the civic domain.

$$OP_{cv} = MotifNet(\mathbf{I_{cv}})$$

The goal of the Discriminator is formulated as a supervised classification training objective:

$$\mathcal{L}_d(\theta_d) = -\sum_{OP_{cv},\ n=1}^{N} \log \mathbf{C}_d(l^n|y^n) \tag{3}$$

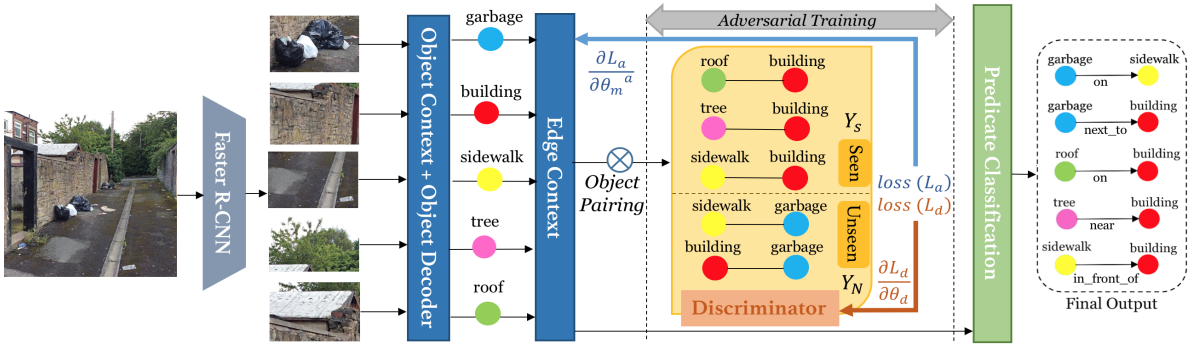

Figure 3: An illustration of our model: Faster R-CNN provides the object labels and their bounding regions. Object context generates a contextualized representation for each object. Edge context generates a contextualized representation for each edge using the representation of the object pairs. During *adversarial training*, information regarding the edge context is passed on to the Discriminator, which learns to distinguish between the *seen* and *unseen* object pairs. The training objective of the Discriminator results in gradients flowing into the Discriminator as well as the edge context layer. The loss for the model decreases as the model learns to fool the Discriminator by adapting a uniform representation for *seen* and *unseen* classes.

$$l^n = \begin{cases} 1(seen) & \text{if } y^n \in Y_s \\ 0(unseen) & \text{if } y^n \in Y_n, \end{cases}$$

where $y^n = (o_i, o_j)^n$, and Y_n and Y_s are the set of object pairs corresponding to R_n and R_s, respectively. θ_d denotes the parameters of the Discriminator to be learned. We minimize $\mathcal{L}_d$ while training the discriminator.

Training Model In accordance with the inverted label loss described above, the training objective of the model is as follows:

$$\mathcal{L}_a(\theta_m^a) = -\sum_{OP_{cv},\, n=1}^{N} \log \mathrm{C}_d(l^n|y^n) \quad \forall y^n \in Y_n$$

Here $\theta_m^a : \{\mathbf{W}_h, \mathbf{W}_t\}$ denotes the parameters of the model that are updated during adversarial training. We minimize $\mathcal{L}_a$ while updating the model.

6 EXPERIMENTS AND EVALUATION

The simplest approach to identify the civic issue from an image is to classify the image into predefined categories. To this end, we trained a classifier using a set of 10 most frequent categories as defined on FixMyStreet forum. On the test dataset, while the accuracies for the three most accurate classes were 86.5%, 83.6% and 75.2%, 4 out of 10 classes had their accuracy less than 17%. Such large variation in accuracies for different classes indicates that classifying images into pre-defined categories is not sufficient. For more details of the classifier, please refer to the extended version of this work [18]. We now provide the implementation details of our model.

6.1 Implementation Details

Discriminator used in adversarial training consists of 3 fully connected layers: two layers with 4096 hidden units followed by the final softmax output. Each hidden layer is followed by a batch normalization, leakyReLU activation function with negative slope of 0.2 and apply a dropout in the training phase with keeping probability of 0.5. Both discriminator and model are trained using ADAM optimizer with a learning rate of 1.2 x 10^{-2} and 1.2 x 10^{-3}, respectively. The value of N_d is set to 150 steps, while N_m is set to 50 steps, with the model and the discriminator being trained iteratively for 12 epochs. For complete implementation details regarding the pre-training of the scene graph model and the Faster R-CNN training, please refer to the extended version of this work [18].

6.2 Evaluation Metrics

Previous work [32] defines three modes for analyzing a scene graph model: 1) Predicate Classification (PredCls) which measures the model's performance for detecting the predicate, given a set of object pairs, in isolation from other factors, 2) Scene Graph Classification (SGCls), which measures the model's performance for predicting the right object labels and predicates, given a set of localized objects, and 3) Scene Graph Generation (SGGen), which requires the model to simultaneously detect the set of objects and predict the right predicate for each object pair. For our approach of generating *CG* using existing scene graph models, it is appropriate to report: (i) the performance of the existing model when generalized to this new domain, and (ii) the accuracy of the output *CG* for representing the civic issue in the image. Deriving from the existing set of tasks, we define a new set of tasks which can help in evaluating our model along these dimensions:

- **OPCls**: the task is to predict the set of object pairs which are indicative of the civic issue present in the image.
- **CGCls**: the task is to predict the set of relations which can represent the civic issue present in the image.
- **CGGen**: the task is to simultaneously detect the region in the image and predict the relations indicative of the civic issue.

For task **OPCls**, we report the experimental results, and use human evaluation for the task **CGCls** and **CGGen**. In accordance with previous work, for **OPCls**, we report results for the image-wise recall metrics (R@k). Since our task is to predict object pairs which are found in civic domains, we report results for *R*@1, *R*@5, *R*@10 & *R*@20 metrics. For **CGCls** and **CGGen**, we report the results using both Precision and Recall metrics ($k : \{1, 3, 5\}$)

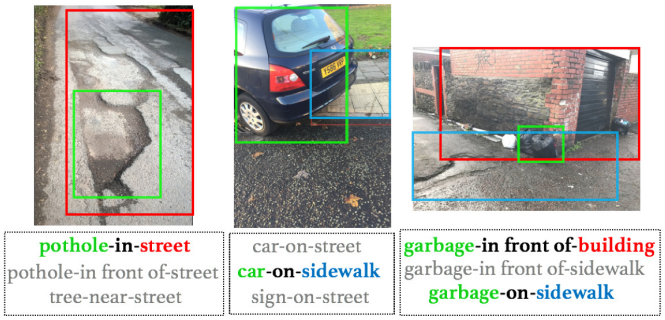

Figure 4: Qualitative examples presenting the Civic Issue Graphs generated by our model. We show the top 3 relations and highlight the ones that are representative of the civic issue along with their bounding regions

6.3 Experimental Results

6.3.1 Removing Object Decoder. The MotifNet model after adversarial training performed poorly when tested on the images from civic domain ($R@10 = 10.5$). We found that the object decoder is not able to predict the correct object labels when the input image contains new objects from the civic domain, as the model has not been trained on these labels. On removing the decoder during test time (denoted as $MotifNet_{Adv}$ in the table), the performance improves significantly ($R@10 = 76.0$, Table 1). Adapting the decoder to a new domain requires ground-truth data in terms of the sequence of objects and the labels, which may not be possible for the civic domain. Therefore, we decided to pre-train the MotifNet model without the decoder (denoted by $MotifNet^{wd}$) and directly use the object labels predicted by the Faster R-CNN. On updating the new model using adversarial training (denoted by $MotifNet^{wd}_{Adv}$), the performance improved significantly, particularly for $R@1$ and $R@5$. Table 1 shows the comparison between the different settings with $MotifNet^{wd}_{Adv}$ performing significantly better than all other models, for all the metrics.

6.3.2 Adversarial Training vs Fine-tuning. The results so far show that using adversarial training can significantly improve the performance of the model. As an alternative approach, we also try to adapt the pre-trained model to our new domain by fine-tuning the predicate classification in the model. Mathematically, we aim to increase the value of $\mathbf{P}(p_{i\to j}|B, O))$, where $(o_i, p_{i\to j}, o_j)$ correspond to R_{CG}. The training objective for this phase is defined as:

$$\mathcal{L}_f(\theta^f_m) = -\sum_{OP_{cv},\ n=1}^{N} l \log \mathbf{P}(p^n_{i\to j}|B^n, O^n) \tag{4}$$

$$l = \begin{cases} 1 & \text{if } p^n_{i\to j} \in R_{CG} \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta^f_m : \{\mathbf{W}_r, \mathbf{w}_{o_i,o_j}\}$, *i.e.*, the weights and bias of the predicate classifier of the model. We minimize L_f while fine-tuning the model which is trained for 6 epochs. Table 1 shows that fine-tuning a pre-trained MotifNet model ($MotifNet^{wd}_{fine}$) brings slight improvement in the performance when compared to the original model ($MotifNet$). However, the model with adversarial training ($MotifNet^{wd}_{Adv}$) performs significantly better than the fine-tuned model ($MotifNet^{wd}_{fine}$). Fine-tuning the model will only improve the detection of relations which are already *seen* by the model, while adversarial training will generalize the performance across both *seen* and *unseen* classes. Further fine-tuning the adversarially updated MotifNet ($MotifNet^{wd}_{Adv+fine}$) model brings no improvement in the performance.

Model Settings	R@1	R@5	R@10	R@20
MotifNet	35.6	64.9	75.4	79.7
$\text{MotifNet}_{\text{Adv}}$	37.7	65.7	76.0	79.8
$\text{MotifNet}^{\text{WD}}$	37.7	63.0	73.3	78.9
$\text{MotifNet}^{\text{WD}}_{\text{Adv}}$	**43.3**	**67.7**	**76.3**	**80.2**
$\text{MotifNet}^{\text{WD}}_{\text{FINE}}$	38.9	63.6	73.8	79.2
$\text{MotifNet}^{\text{WD}}_{\text{Adv+FINE}}$	43.1	67.7	76.3	80.2

Table 1: Recall for different settings; *Adv*: Adversarial Training; *fine*: fine-tuning; *wd*: *without decoder* setting

6.4 Human Evaluation

To establish the efficacy of our model at appropriately representing civic issues from images, we recruited Amazon Mechanical Turk workers. We randomly sampled 300 images from the test set; each image was evaluated by 3 workers. In accordance with our definition of *CG*, we retained only the top 5 relations from the output relations for which $o_i \in O_1$ and $o_j \in O_2$, where (o_i, o_j) denotes the unordered pair of objects in a relation.

The evaluation was carried out for the tasks **CGCLS** and **CGGEN** in two phases. For **CGCLS**, workers were shown an image along with 5 relations and were asked to select 0 or more relations that appropriately represented the civic issue(s) in that image. An option was also given to specify any additional relations separately. For **CGGEN**, we retrieved the relations which were marked as relevant for a given image. For each such relation, workers were shown the bounding regions for the objects present in the relation, and asked to evaluate the coverage of these bounding regions, on a scale of 0-10. We report two metrics – Precision and Recall, for both the tasks and consider only the majority voted relations with a minimum average rating of 5 for the bounding regions. Table 2 shows the performance of our model. The results show that 83.3% of the times, the relation representing a civic issue is present in the top 3 relations of our *CG*, and 53.0% of the times, the top relation itself represents a civic issue in the image. The accuracy on the **CGGEN** task further indicates that our model is capable of generating accurate groundings for the objects representing the civic issue.

7 CONCLUSION

We introduce a novel unsupervised mechanism of adapting existing scene graph models via adversarial training and present an application of our approach for generating Civic Issue Graph. Our approach can be utilized in existing platforms, which allow users to report civic issues using images. Once the user uploads an image,

	CGCLS			CGGEN		
	@1	@3	@5	@1	@3	@5
Precision	53.0	31.9	24.7	50.9	30.3	24.1
Recall	53.0	84.0	99.0	50.9	83.3	99.0

Table 2: Precision and Recall values for the tasks CGCLS and CGGEN based on human evaluation

our model can automatically generate text-based relations (*e.g.,* garbage-on-street, garbage-next to-building) depicting the civic issue in the input image. These text-based relational descriptions can be shared with the authorities, which can be utilized for large scale analysis, thereby automating the process and removing any dependency on the actual image uploaded by the user. Furthermore, if needed, natural language descriptions can be generated from these relations using a template-based approach. Our experimental analysis helps provide a framework for adapting scene graph models to other settings as well. We also release two multi-modal (text and images) datasets with information of civic issues, to encourage future work in this domain.

REFERENCES

[1] 2016. FixMyStreet. (2016). https://www.fixmystreet.com/

[2] 2017. Mayor's Management Report. https://www1.nyc.gov/assets/operations/downloads/pdf/mmr2017/2017_mmr.pdf. (2017).

[3] Deborah Agostino. 2013. Using social media to engage citizens: A study of Italian municipalities. *Public Relations Review* 39, 3 (2013), 232–234.

[4] Shubham Atreja, Pooja Aggarwal, Prateeti Mohapatra, Amol Dumrewal, Anwesh Basu, and Gargi B Dasgupta. 2018. Citicafe: An Interactive Interface for Citizen Engagement. In *23rd International Conference on Intelligent User Interfaces*. ACM, 617–628.

[5] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, Adapt and Tell: Adversarial Training of Cross-Domain Image Captioner.. In *ICCV*. 521–530.

[6] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, Adapt and Tell: Adversarial Training of Cross-Domain Image Captioner.. In *ICCV*. 521–530.

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In *Computer Vision and Pattern Recognition (CVPR)*.

[8] Annalisa Cocchia. 2014. Smart and digital city: A systematic literature review. In *Smart city*. Springer, 13–43.

[9] Peter Dahlgren. 2011. Parameters of online participation: Conceptualising civic contingencies. *Communication management quarterly* 21, 4 (2011), 87–110.

[10] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15)*. JMLR.org, 1180–1189. http://dl.acm.org/citation.cfm?id=3045118.3045244

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[12] Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, Vol. 1. 44–52.

[13] IChangeMyCity. 2012. https://ichangemycity.com. (2012).

[14] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3668–3678.

[15] Maria Karakiza. 2015. The impact of social media in the public sector. *Procedia-Social and Behavioral Sciences* 175 (2015), 384–392.

[16] Matthew Klawonn and Eric Heim. 2018. Generating Triples with Adversarial Networks for Scene Graph Construction. *arXiv preprint arXiv:1802.02598* (2018).

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[18] Shanu Kumar, Shubham Atreja, Anjali Singh, and Mohit Jain. 2019. Adversarial Adaptation of Scene Graph Models for Understanding Civic Issues. *arXiv preprint arXiv:1901.10124* (2019).

[19] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. 2018. Visual Relationship Detection with Deep Structural Ranking. (2018).

[20] Hiroya Maeda, Yoshihide Sekimoto, Toshikazu Seto, Takehiro Kashiyama, and Hiroshi Omata. 2018. Road Damage Detection Using Deep Neural Networks with Images Captured Through a Smartphone. *arXiv preprint arXiv:1801.09454* (2018).

[21] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[22] Graeme Mearns, Rebecca Simmonds, Ranald Richardson, Mark Turner, Paul Watson, and Paolo Missier. 2014. Tweet my street: a cross-disciplinary collaboration for the analysis of local twitter data. *Future Internet* 6, 2 (2014), 378–396.

[23] Ines Mergel. 2012. Distributed democracy: Seeclickfix. com for crowdsourced issue reporting. (2012).

[24] Gaurav Mittal, Kaushal B Yagnik, Mohit Garg, and Narayanan C Krishnan. 2016. Spotgarbage: smartphone app to detect garbage using deep learning. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 940–945.

[25] Nitish Mittal, Swati Agarwal, and Ashish Sureka. 2016. Got a Complaint?-Keep Calm and Tweet It!. In *International Conference on Advanced Data Mining and Applications*. Springer, 619–635.

[26] Taewoo Nam and Theresa A Pardo. 2011. Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times*. ACM, 282–291.

[27] Paolo Neirotti, Alberto De Marco, Anna Corinna Cagliano, Giulio Mangano, and Francesco Scorrano. 2014. Current trends in Smart City initiatives: Some stylised facts. *Cities* 38 (2014), 25–36.

[28] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1717–1724.

[29] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-Adversarial Domain Adaptation. In *AAAI*.

[30] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*. 70–80.

[31] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 4.

[32] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2.

[33] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural Motifs: Scene Graph Parsing with Global Context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5831–5840.

[34] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian D Reid, and Anton van den Hengel. 2018. HCVRD: A Benchmark for Large-Scale Human-Centered Visual Relationship Detection.. In *AAAI*.