ATTENTIONAL ROAD SAFETY NETWORKS

Sonu Gupta[⋆] *Deepak Srivatsav*[⋆] *A V Subramanyam*[⋆] *Ponnurangam Kumaraguru*[⋆†]

[⋆]IIIT-Delhi, [†]IIIT-Hyderabad

ABSTRACT

Road safety mapping using satellite images is a cost-effective but a challenging problem for smart city planning. The scarcity of labeled data, misalignment and ambiguity makes it hard to learn efficient embeddings in order to classify between safe and dangerous road segments. In this paper, we address the challenges using a region guided attention network. In our model, we extract global features from a base network and augment it with local features obtained using the region guided attention network. In addition, we perform domain adaptation for unlabeled target data. In order to bridge the gap between safe samples and dangerous samples from source and target respectively, we propose a loss function based on within and between class covariance matrices. We conduct experiments on a public dataset of London to show that the algorithm achieves significant results with the classification accuracy of 86.21%. We obtain an increase of 4% accuracy for NYC using domain adaptation network.

Index Terms— Road safety, Region guided network, Domain Adaptation

1. INTRODUCTION

Road accidents remain one of the pressing communal welfare concerns. Regardless of notable advancements in the field of vehicle technology and road engineering, on a global scale, traffic accidents are one of the leading causes of premature death and injury. According to the World Health Organization, more than 1.25M people die every year due to traffic accidents [1]. Therefore, minimizing the road accidents is a worldwide challenge and can benefit a majority of the nations in different ways. Towards this, it is important to understand which road segments are potentially dangerous or safe.

Few works have shown the influence of environmental factors like weather, light condition on road accidents [2, 3, 4]. Due to lack of resources and technology, such data is not maintained properly in most low and middle-income nations and unfortunately, these are the nations which suffer dreadfully from traffic accidents [5]. Hence, there is a need for an efficient approach which can work well with easily available and affordable data. To this end, satellite images are used for road safety mapping [6]. However, the dataset is mostly imbalanced, and percentage of the safe class is far more compared to the dangerous class. Further, the images are misaligned. Thus, using such data for training efficient supervised models is another challenge. Besides, it has been observed that models trained using classical machine learning techniques for one region do not perform well if tested on regions that differ immensely in terms of traffic regulations, city planning, architecture, etc. [6]. Also, it is quite inconvenient to obtain traffic accident data for every different region and train a model for the same. Hence, the difference in domains is yet another challenge in the effective road safety mapping.

To address these challenges, we propose a deep region guided attention network. We use a base network (ResNet-50) to extract global features. We use a sub-network which can attend to individual subregions. Towards this, we extract the conv 2 layer features and divide them into N non-overlapping regions. Region or part based networks have shown good accuracy in re-identification tasks [7]. Further, in case of unlabeled target data, we augment this network to adapt to different domains. We use a loss function based on the covariance matrix to minimize the gap between the source and target domains. Our contributions are: (a) we propose a deep learning framework that uses the region guided attention network to predict city-scale safety maps from satellite images, and (b) we propose a domain adaptation network with a training loss which minimizes the Frobenius norm of the difference of *within* class covariance matrices of source and target, as well as the difference of *between* class covariance matrices of source and target.

2. RELATED WORK

In the recent past, the remarkable progress in deep learning has contributed significantly to the field of computer vision. Spatio-temporal data have been used by researchers to predict the number of accidents in a given area with the help of ConvLSTM [8]. [9] developed a Stacked Denoising Autoencoder for prediction of traffic accident risk level at the city-scale using real-time GPS data of users. In a similar study, [6] demonstrate that visual attributes captured in a satellite image can be used as a proxy signal of road safety. They proposed a deep-

learning based mapping framework that exploits open data to predict city-scale safety maps at bearable costs. However, the distinction among the above frameworks resides in their application itself. Both [9, 8] are interested in the real-time prediction of traffic accidents, whereas [6] is inclined towards assisting in informed decision-making for city-planning and policy formulation when heterogeneous data is not accessible or bearable. Our work is closer in spirit to [6]; however, there are three major differences. First, we only consider safe and dangerous classes compared to an additional neutral class defined by [6]. This is because a neutral road segment may not assert a strong decision making during road safety planning. Second, the authors extract features from AlexNet, whereas, we propose a novel architecture. Third, we also perform domain adaptation for the unlabeled target domain.

3. OVERVIEW

We select two cities, London (UK) and New York City (USA) for this study. We select these two due to availability of sufficient data and the extreme difference in traffic regulations, city planning, architecture etc. All the data that we use in this work are available as *open data*. We also present results for Denver (USA) which is altogether a different domain.

Traffic Accident Data: We collected 99,516 traffic accident records for London from 2013 to 2016. We gathered 1,256,205 traffic accident reports for New York City (NYC) from 2012 to 2017. Similarly, we collected 143,776 traffic accident reports for Denver from 2013 to 2017.

Problem Formulation: We divide the whole study area by imposing a grid G, where each grid g_i is a square region of $s \times s$. Here, s is equal to 30 meters which is a sufficient area for traffic accident analysis [6]. In this work, our objective is to produce a model that learns features such that it can predict the safety level of a given grid g_i using its raw satellite image. In order to curate a training dataset, we plot each accident on the grid using the locations from the Traffic Accident data. Each grid g_i is given a safety score S_i which is equal to the number of accidents in the grid. The higher S_i corresponds to a higher number of accidents. We obtain a satellite image for each of these grids. To obtain the image label for each g_i, we apply the classical k-means algorithm to bin S_i into two clusters. Hence, we get two clusters of *safe* and *dangerous* locations. We observe that the obtained set of labels are highly imbalanced. Over 88% of the regions are labeled as safe. Therefore, we re-sample the dataset by down-sampling the majority class such that both classes are balanced out.

4. DEEP ATTENTION-NETWORK APPROACH

In this section, we discuss the network architecture and the implementation of the domain adaptation.

Network Architecture: The model consists of a global and a local network, with the global network borrowing layers from ResNet50 [10]. We show the architecture of the proposed model in Figure 1. There are three 2-way switches in

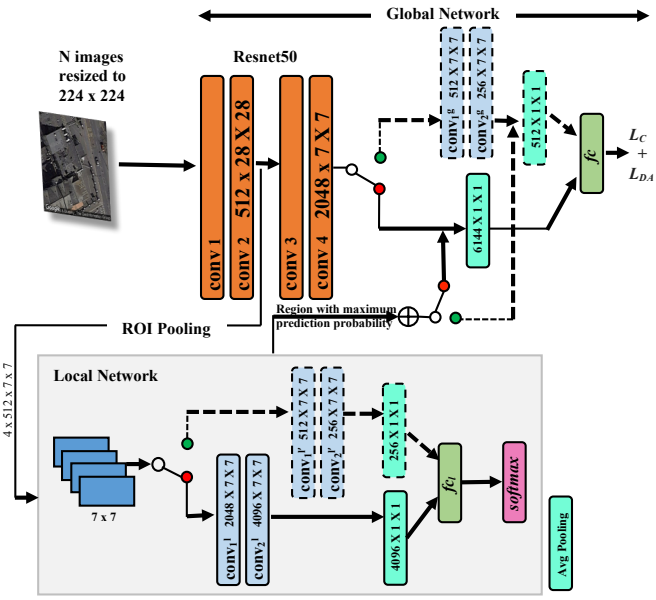


Fig. 1: Architecture of our proposed network. (Best viewed in color).

the architecture which enables the network to run in two different modes. When all the switches are connected to *red* nodes, the Deep Attention Model (DAM) mode is enabled whereas if switches are connected to all the green nodes, the Deep Attention Model with Domain Adaptation (DAM-DA) mode is enabled. The solid and dashed lines represent DAM and DAM-DA mode respectively. We discuss the DAM-DA mode under Domain Adaptation subsection. As shown in Figure 1, the global network is augmented with a local network. In our experiments, we notice that the global network based on ResNet only, gives significant false positives, where false positives are defined as dangerous regions being predicted as safe. A possible reason behind such false positives could be that the satellite images are usually misaligned. To overcome misalignment and improve the performance of ResNet, we design the network such that it can exploit local regions in addition to the global region. We perform *Region of Interest* (ROI) pooling on conv 2 layer of the model with the subregions being N non-overlapping blocks. We pass features of each subregion through a local network. The local network consists of conv_1^l and conv_2^l layers. Each subregion of size 7x7 acts as an input to the local network to extract features and to identify which of them is useful for the classification task. Thus, this network attends to the regions which can significantly contribute towards enhancing the classification accuracy. We perform average pooling on the output of conv_2^l layer. We obtain a feature vector of size 4096x1x1 that is passed through fc_l. The subregion with the maximum prediction probability of either of the two classes (Safe and Dangerous) is concatenated to the feature-map of the global network. We perform concatenation after the conv 4 layer of the global network. Thus, we can see that only certain subregions can guide the global feature vector for classification. Once aver-

age pooling is performed, we pass the feature vector through fc to obtain the final d-dimensional vector. We train the network using cross entropy loss (L_C).

Domain Adaptation: Now, we discuss our proposed Domain Adaptation (DA) technique for the target domain where there is no annotated data available. In our experiments, we consider London as the source domain for which labeled data is present, whereas, NYC as the target domain for which labels are not given. There are significant differences in both the cities. Although NYC is nearly 20% smaller in terms of area, it's population density is almost double than that of London. According to [11], higher the population density, higher are the chances of occurrence of accidents. Due to abovementioned reasons, these cities would be a good extreme example for performing DA and can give us an idea of lower bound of model's performance. First, we train DAM network using data from the source domain. Using this trained model, we generate *pseudo labels* for data from the target domain. Thus, we have annotated labels for the source domain (London) and pseudo labels for the target domain (NYC). Now, we use data from both source and target domain to train the augmented DA (DAM-DA) network as shown in Figure 1. There are three differences when compared to DAM. First, this network has two additional convolutional layers in the global network, conv_1^g and conv_2^g layers of dimensions 512x7x7 and 256x7x7 respectively, to reduce the dimensions of the feature maps. This augmentation is necessary to be able to compute the covariance matrix efficiently. We explain the utility of the covariance matrix later in this section. Second, instead of conv_1^l and conv_2^l layers in the local network, we use $\text{conv}_1^{l'}$ and $\text{conv}_2^{l'}$ layer of size 512x7x7 and 256x7x7 respectively. This is also done to reduce the dimensions for covariance matrix calculation. Third, in DAM mode, we train the network using L_C only, but in DAM-DA mode we use domain adaptation loss (L_{DA}) along with L_C. We explain L_{DA} next.

Let x_i denote the feature for *i*-th sample classified as dangerous in source domain. Similarly, let y_i be the feature for *i*-th sample classified as safe in source domain. These features are obtained from the final layer (fc) of the network as shown in Figure 1. Now, we can obtain the *within* class covariance matrix $\sum_{SW} \in \mathcal{R}^{d\times d}$ as,

$$\sum_{SW} = \sum_{i,j\neq i}(x_i - x_j)(x_i - x_j)^T + \sum_{i,j\neq i}(y_i - y_j)(y_i - y_j)^T \quad (1)$$

Similarly, we can compute the *between* class covariance matrix $\sum_{SB} \in \mathcal{R}^{d\times d}$ as,

$$\sum_{SB} = \sum_{i,j}(x_i - y_j)(x_i - y_j)^T \quad (2)$$

We can compute the *within* and *between* class covariance matrices for target domain in a similar manner. Let these be denoted by $\sum_{TW}$ and $\sum_{TB}$ respectively. Then, we use the following loss function to adapt to the target domain,

$$L_{DA} = \left\|\sum_{SW} - \sum_{TW}\right\|_F^2 + \left\|\sum_{SB} - \sum_{TB}\right\|_F^2 \quad (3)$$

where $\|.\|_F$ denotes the Frobenius norm. Here, instead of coarsely treating London and NYC as source and target datasets, we use the knowledge of available labels. The first component bridges the gap between domains of *within* class between source and target, while the second component bridges the gap for *between* class. Though the pseudo labels are noisy, it is still beneficial to use them. In our experiments, we demonstrate that using the loss to minimize the distance between feature covariance matrices between source and target, which is agnostic of source labels, achieves sub-par performance when compared to L_{DA}.

5. EXPERIMENTS

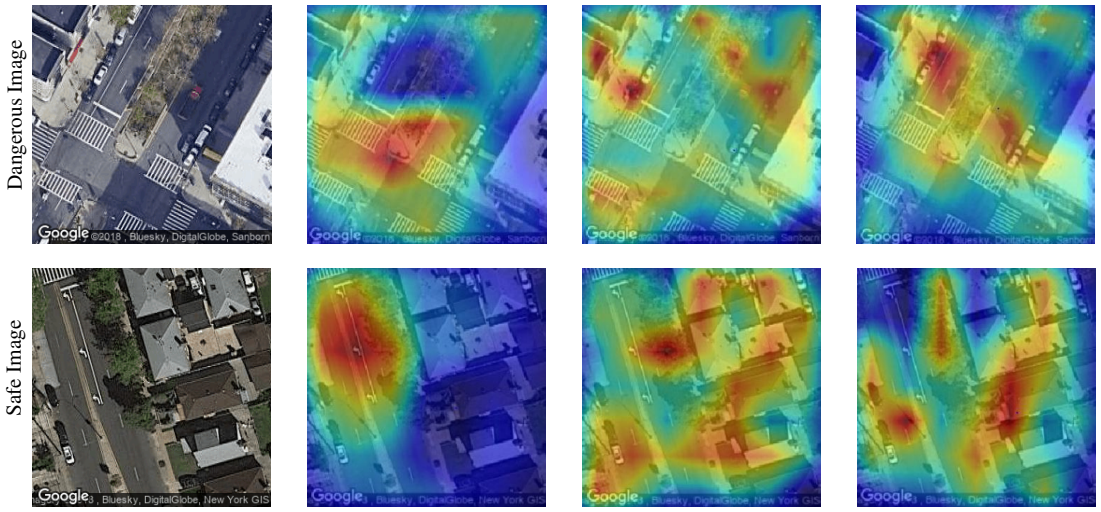
We initially performed the experiments with AlexNet, VggNet, DenseNet, and ResNet and found that ResNet gives the best accuracy. First, we train our network on London dataset. There are 4,517 training samples and 903 validation samples in each class. The models are trained with a batch size of 4, with learning rate as 0.0001 and a learning rate decay of 0.5 per 10 epochs. We train it over 50 epochs. To assess the model, we test on unseen data from London. As shown in Table 1, we test the models in multiple scenarios such as with horizontal (DAM + HS), vertical (DAM + VS), and square boxed (DAM + SQ) subregions and with the combinations of them. We test another variation of the model where the activation maps are produced based on an intermediate convolutional layer. After experiments, we decide to use conv 2 layer as we achieve better results with it. We find that the DAM using HS+VS+SQ subregions outperformed ResNet50, VGG19, and its other variants on the same dataset.

Table 1: Comparison of classification accuracy (in %) for ResNet50, VGG19, and variants of the DAM both for original dataset (London) and cross dataset (NYC and Denver).

Model	Original Data	Cross Data	
	London	NYC	Denver
ResNet50	85.77	69.16	70.00
VGG19	85.83	64.60	70.00
DAM (HS)	85.81	72.28	**76.20**
DAM (VS)	85.52	**74.77**	75.00
DAM (SQ)	85.86	70.70	70.00
DAM (HS+VS)	85.34	70.37	70.01
DAM (HS+VS+SQ)	**86.21**	67.23	69.86

Cross Dataset Testing: We also perform cross-data testing and test our model on NYC and Denver dataset. From Table 1, we can see that DAM + VS and DAM + HS performs the best for NYC and Denver dataset respectively. For DAM (HS+VS+SQ), the base model trained on London with three sub-regions is over-fitting, and the network does not generalize well on Denver. When we decrease sub-regions to 2, the accuracy increases for NYC and Denver, and the model performs the best when used with only one sub-region.

Performance Measure: One of our goals is to minimize the false positive (FP). This is necessary as the cost of predicting safe as dangerous may only lead to overly conservative approaches for planning, whereas, having high FP rate with more dangerous locations being marked safe can lead to fatal accidents. Therefore, we consider the classification of dan-



(a) Original Images (b) ResNet50 (c) DAM-Global Net (d) DAM-local Net

Fig. 2: A comparison of activated features of a *dangerous* and a *safe* location among ResNet50, DAM-Global Network, and DAM-Local Network. Top row and bottom row represent dangerous and safe locations respectively. (Best viewed in color).

gerous images as safe to be more costly than vice-versa. We test ResNet and DAM on London, NYC, and Denver test-set of 7,228, 8,342, and 500 images respectively. From Table 2, we can see that DAM gives fewer FP in comparison to ResNet for every domain.

Table 2: A comparison of false positive rate between DAM and ResNet50. All results are in percentage.

Model	London	NYC	Denver
DAM	**07.60**	**22.05**	**29.20**
ResNet50	12.73	39.05	40.80

Qualitative Analysis: In order to understand the behavior of the network qualitatively, we generate Class Activation Maps (CAM) [12]. In Figure 2, the images in the top row correspond to a dangerous image and its CAMs for ResNet50, DAM-Global Network, and DAM-Local Network. Similarly, the images in the bottom row correspond to a safe image and its corresponding CAMs. From the top row in Figure 2c and Figure 2d, we can establish that the DAM not only identifies the green region around the divider on the road but also identifies the cars and roads in the surrounding areas as well, which positively contributes towards the identification of this image as dangerous. This image contains a dangerous location, but, ResNet50 mis-classifies it as safe. As from top row in Figure 2b, ResNet50 seems to have identified the green region and used that to classify the sample as safe, whereas our proposed DAM is efficient enough to correctly identify it as a dangerous location. Similarly, from the CAM in the bottom row, we observe that ResNet50 recognizes only roads whereas the DAM identifies houses and trees in the image. Therefore, DAM correctly identifies this sample as safe whereas ResNet50 mis-classifies it as dangerous.

Evaluation of DAM-DA Network: We train our DAM-DA network with 2,085 images of each class, i.e., Safe and Dangerous. We use 4,170 images each from the source (London) and target (NYC) domain. We use a batch size of 16, a learning rate of 0.0001 and a decay of 0.5 per 10 epochs. It is trained for over 50 epochs. We test the DAM-DA network on 3,336 images from NYC. As shown in Table 3, DAM-DA network gives an accuracy of 75.75% in comparison to DAM network which gives an accuracy of 71.94%. In another setting, we replace L_{DA} with the loss [13] in Equation 4,

$$L_{DA}^{ST} = \frac{1}{d} \|C_S - C_T\|_F^2 \tag{4}$$

where C_S and C_T denotes the feature covariance matrix of source and target samples in a training batch, respectively. When compared to L_{DA}, L_{DA}^{ST} does not take the class labels into account. As shown in Table 3, we obtain an accuracy of 74.73%. Similarly, DAM$-$DA$-L_{DA}$ results in the least false positives rate (18.94%) in comparison to its counterparts. Thus, we can see that the class wise covariance loss works better in this scenario .

Table 3: A comparison of DAM, DAM-DA with L_{DA}, and DAM-DA with L_{DA}^{ST} network with batch-size of 16.

Model	Accuracy (%)	FPR(%)
DAM	71.94	30.03
DAM$-$DA$-L_{DA}$	**75.75**	**18.94**
DAM$-$DA$-L_{DA}^{ST}$	74.73	30.69

6. CONCLUSION

In this paper, we address the challenge of learning efficient embeddings to classify the road segments as dangerous or safe using easily available and inexpensive data. We leverage open data and satellite images to predict city-scale road safety maps. We propose a deep learning based model that uses a region guided attention network. It consists of a global and a local network. The local network attends to features in the subregions and the features with maximum prediction

score are used to guide the global features to enhance the accuracy. We evaluate our network on the public dataset of London and achieve the accuracy of 86.21%. We experiment with the cross-datasets of NYC and Denver and achieve significant results with the accuracy 74.77% and 76.20% respectively. In addition, we propose a covariance loss based domain adaptation for the scenario where target domain labels are missing. In our experiments, we show that with the domain adaptation network, the accuracy of NYC increases by 4% and the network also achieves the lowest false positives.

7. REFERENCES

[1] World Health Organization, *Global status report on road safety 2015*, World Health Organization, 2015.

[2] JD Tamerius, X Zhou, R Mantilla, and T Greenfield-Huitt, "Precipitation effects on motor vehicle crashes vary by space, time, and environmental conditions," *Weather, Climate, and Society*, vol. 8, no. 4, pp. 399–407, 2016.

[3] Younshik Chung, Seonjung Kim, and Seunghoon Cheon, "A framework for modelling crash likelihood information under rainy weather conditions," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2018, pp. 823–832.

[4] Alan W Black and Gabriele Villarini, "Effects of methodological decisions on rainfall-related crash relative risk estimates," *Accident Analysis & Prevention*, 2018.

[5] Uli Schmucker, J Seifert, Dirk Stengel, Gerrit Matthes, Caspar Ottersbach, and Axel Ekkernkamp, "Road traffic crashes in developing countries," *Der Unfallchirurg*, vol. 113, no. 5, pp. 373–377, 2010.

[6] Alameen Najjar, Shun'ichi Kaneko, and Yoshikazu Miyanaga, "Combining satellite imagery and open data to map road safety.," in *AAAI*, 2017, pp. 4524–4530.

[7] Fuqing Zhu, Xiangwei Kong, Liang Zheng, Haiyan Fu, and Qi Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4806–4817, 2017.

[8] Zhuoning Yuan, Xun Zhou, and Tianbao Yang, "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 984–992.

[9] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki, "Learning deep representation from big and heterogeneous data for traffic accident inference.," in *AAAI*, 2016, pp. 338–344.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Seppo Nieminen, Olli-Pekka Lehtonen, and Miika Linna, "Population density and occurrence of accidents in finland," *Prehospital and disaster medicine*, vol. 17, no. 4, pp. 206–208, 2002.

[12] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[13] Baochen Sun and Kate Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 443–450.

[14] Sonu Gupta, Deepak Srivatsav, AV Subramanyam, and Ponnurangam Kumaraguru, "Attentional road safety networks," *arXiv preprint arXiv:1812.04860*, 2018.