Article

A Ensemble Deep Learning Method for Vehicle Type Classification on Visual Traffic Surveillance Sensors

Wei Liu [1,4], Miaohui Zhang [2,4,*] Zhiming Luo [3] and Yuanzheng Cai [5]

1 Virtual Reality and Interactive Techniques Institute, East China Jiaotong University, Jiangxi, China
2 Institute of Energy, Jiangxi Academy of Sciences, Jiangxi, China
3 Cognitive Science Department, Xiamen University, China
4 Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, Yi Chang 443002,China
5 Department of Computer Science, Minjiang University, Fujian, China
* Corresponding author.

Abstract: Visual traffic surveillance systems play important roles in intelligent transport systems nowadays. A visual traffic surveillance system usually needs to correctly detect objects from images or videos and classify them into different categories *(e.g. car, truck, bus)*. This paper aims to introduce a new vehicle type classification scheme on the images pictured by multi-view visual traffic surveillance sensors. Most image classification algorithms focus on maximizing the percentage of the correct predictions, which have a deficiency that images from minority categories are prone to be misclassified as the dominant categories. To address this challenge of classifying imbalanced data acquired from visual traffic surveillance sensors, we propose a method which integrates deep neural networks with balanced sampling in this paper. The proposed method consists of two main stages. In the first stage, balanced sampling is applied to alleviate the unbalanced data set problem. In the second stage, an ensemble of convolutional neural network models with different architectures is constructed with parameters learned on the augmented training data set. Experiments on the MIO-TCD classification challenge dataset demonstrate that the proposed method is able to improve the performance compared with the baseline algorithms.

Keywords: traffic data; traffic surveillance systems; intelligent transport systems; image classification; ensemble learning; imbalanced data

0. Introduction

In the last decade, we have seen a worldwide rise of using visual traffic surveillance systems, due to the rapidly growth of storage power, computation speed and the innovations in video compression standards. For the first step, a visual traffic surveillance system usually needs to correctly detect objects from images or videos and classify them into different categories *(e.g. car, truck, bus)*. Efficient and robust classification can lead to many semantic results, such as "pedestrian no.1 is moving, car no.3 stopped" or some more advanced results such as "van no.8 is turning right, bicycle no.5 is moving at a speed of 10 kilometers per hour." However, such high-level information is possible only if we can correctly detect and classify the objects .

With the increasing amount of available data, image processing has emerged to be a hot spot in the field of artificial intelligence and image classification is one of fundamental tasks. As is shown in Figure 1, the goal of image classification is to assign a predefined category label to an image. Image classification has a wide application in the field of artificial intelligence, including self-driving, augment

 doi:10.20944/preprints201709.0037.v1

reality, etc [1–3]. Recently, image classification have attracted more and more research interest. Though image classification has been widely studied in the academia and deployed in the industry, it is is not a trivial task, still a challenging task. For example, many practical image classification data are imbalanced, i.e., some of the categories are represented by only a few samples, while some others make up the majority.

In the field of traffic surveillance, a visual traffic surveillance system needs to detect vehicles or pedestrains and classify them if possible. In the practical application, *Pedestrains*, *Bicycles* and *Motorcycles* often make up minority of the data set, in contrast with *Cars* and *Buses*. Consequently, to avoid the misclassification of images from majority categories as rare classes, it is also not appropriate to assume misclassification errors cost for all samples are equal. If misclassification errors cost are implicitly assumed to be equal, images from minority categories are prone to be misclassified to be the dominant categories. However, image classification plays an important part in visual intelligent transport systems. It is a prerequisite for semantic results of visual traffic surveillance systems. Therefore, to effectively reduce the number of fatalities, it is reasonable to focus on enhancing the mean precision of all categories, in the condition of high overall accuracy.

In the field of machine learning, learning from imbalanced data has been studied actively for about two decades. It's been the subject of many papers, workshops, special sessions, and dissertations [4–7,7–9] and data manipulation techniques [10–14]. However, there is no definite answer for *What is the best machine learning algorithm for imbalanced data classification?* it depends on the data. The approaches to tackle the problem of extremely imbalanced data can be mainly generalized into two main kinds. One is based on cost sensitive learning[15]. At first, misclassification of the minority class is assigned a high cost. Then, try to minimize the overall train error. The other way is to employ a sampling tactic. Most research has been focused on this approach based on balanced sampling. There are three common sampling approaches, including oversample the minority class, undersampling the majority class, and Synthesizing new minority classes. The easiest approaches to balance the training set are oversampling and undersampling, which require little change to the processing steps, and simply involve adjusting the example sets until the example classes contained in the training set are balanced. Oversampling randomly replicates minority instances to increase their population. Learning from imbalanced classes continues to be an ongoing area of research in machine learning with new algorithms introduced every year.

To tackle the imbalanced problem for traffic data acquired from visual traffic surveillance sensors, we propose an CNN-based deep learning framework to increase mean precision in this paper. We focus on integrating deep neural networks with balanced sampling. As is shown in Fig.2, the proposed approach consists of two stages. In the first stage, balanced sampling is applied to alleviate the unbalanced data set problem. In the second stage, an ensemble of convolutional neural network models with different architectures is constructed with 8 parameters learned on the augmented training data set.

The outline of this paper is organized as follows. Section 1 surveys related work . The detailed of the proposed method is presented in Section 2. Experimental results and comparison are provided in Section 3. Finally, the conclusion of this paper is in Section 4.

1. Related work

1.1. Imbalanced Data Classification

In recent years, there has a spate of interest in learning from imbalanced data in data mining and machine learning. A vast number of techniques have been tried, and the previous work to tackle the class imbalanced problem can be mainly categorized into the algorithm oriented approaches [4–9,16] and data manipulation techniques [10–14]. The former category aims to study and modify the training algorithms to achieve better performance in imbalanced data classification, by adjusting

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2017 doi:10.20944/preprints201709.0037.v1

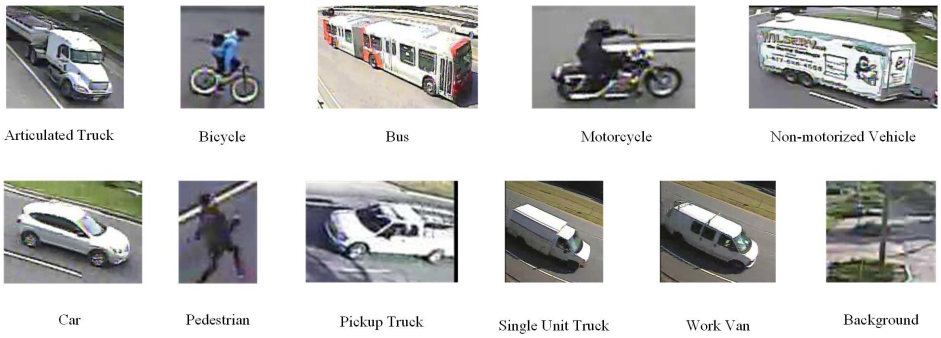

Figure 1. MIO-TCD classification challenge dataset acquired from visual traffic surveillance sensors.

misclassification costs. The latter category, instead of manipulating samples at the algorithmic level, operates at the data level by data re-sampling. A comprehensive review is presented in [13].

There are several types of data manipulation techniques, which can be mainly divided into two groups: oversampling and undersampling. In the last two decades, data manipulation techniques based on oversampling and undersampling have been widely studied to counter the effect of imbalanced data sets[17]. Different variants of oversampling and undersampling techniques have been tested for imbalanced data sets in the literature [18]. In oversampling, minority instances are generated by certain algorithms to make the data set balanced. The easiest approaches for resampling is to randomly replicates minority instances to increase their population, or to randomly downsample the majority class. The positive consequence for replication-based random oversampling is that it duplicates the number of errors for minority instances: if minority instances on the data set is replicated four times, the classifier will make five errors on the new set. Although oversampling results in more data, it is not superior versus undersampling which throwing away data. Replicating data is not without consequence, since it results in duplicate data, which makes variables appear to have lower variance than they do. That is to say, replication-based random oversampling has a tendency to overfit, because it does not actually increase any information actually. To address this, Chawla et al. [10] proposed synthetic minority over-sampling technique(SMOTE), generating new non-replicated minority examples. Several improved variants of SMOTE are presented in [12,13,19]. However, these method can potentially lead to overfitting, since their broaden decision regions are still error-prone by synthesizing noisy and borderline instances. On the other hand, downsampling is to throw away a part of majority samples to make the data set balanced. Downsampling is very efficient, since the new data set is only a subset of the imbalanced data set. The main disadvantage is that potentially valuable information may be removed by throwing away a part of the majority samples. However, undersampling is often preferred to oversampling [11], although it may result in loss of information.

To avoid the disadvantages of re-sampling based techniques, there are many studies focusing on algorithm oriented approaches. Liu et al. proposed two algorithms named *EasyEnsemble* and *BalanceCascade* respectively [20]. Sun et al.[21] proposed a taxonomy to organize all proposed strategies following the training and the test phases in text classification tasks, using Support Vector Machines. *EasyEnsemble* samples some subsets from the dominant classes, trains a classifier using each of the subset, and then combine the outputs of the trained classifiers. *BalanceCascade* trains the models sequentially, and removes correctly classified major class examples of the trained classifiers from further consideration in each step of the training. Tang et.al [9] make classic SVM cost-sensitive to improve classification on highly skewed datasets. Zadrozny [4] proposed a family of methods for converting classifier learning algorithms and classification theory into cost-sensitive algorithms and theory, based on cost-proportionate weighting of the training examples. This method achieves better predictive performance, while drastically reducing the computation required by other baseline methods. To combat imbalance, Ting [16] studied how to improve our understanding of various cost-sensitive boosting algorithms and how variations in the boosting procedure affect misclassification cost and

 doi:10.20944/preprints201709.0037.v1

high cost error. Chen et al. [8] proposed two methods to to overcome the imbalanced data classification problem utilizing random forest. One is based on cost sensitive learning, and the other is based on a sampling technique. The two proposed methods are less vulnerable to noise than boosting.

1.2. Deep Learning

So far, we have given a brief summary of the classification techniques in the field of tackling the imbalanced problem. However, there are many researches on classification in other fields deserving our attention in the literature. In the last decade, deep neural networks have led to a series of breakthrough results on a variety of machine learning tasks, such as computer vision, text analysis and voice recognition, etc. One of the essential components bringing about these breakthrough results has been a special kind of neural network architecture called a convolutional neural network (CNN), which can be thought of as a kind of neural network that uses many identical copies of the same neuron. Since AlexNet was proposed by Krizhevsky et al [22], deep learning methods have shown superior performance for image classification, compared with conventional "shallow learning" models. AlexNet has been successfully applied in a variety of compute vision tasks, such as object detection [23], video classification[24], and segmentation [25], etc. These successes spurred a new line of research that focused on devising higher performance convolutional neural networks. The quality of network architectures has been significantly improved by utilizing deeper and wider networks, since 2014. Lin et al. proposed Network-in-Network [26], indicating how the introduction of residual connections leads to dramatically improved training speed for the Inception architecture. Simonyan et al. proposed VGGNet [27], facilitating further research on the use of deep visual representations in computer vision. Szegedy et al. presented GoogLeNet codenamed Inception [28], setting the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 [1]. However, deeper neural networks are more difficult to train. When deeper networks are able to start converging, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly. To tackle with the degradation problem in the stage of training for very deep neural networks, He et al. presented a residual learning framework named ResNet [29] that are substantially deeper than those employed previously. Their submissions based on ResNets won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation in the ILSVRC & COCO 2015 competitions[2].

Although deep learning has been shown as a successful machine learning method for a variety of tasks, to our best knowledge, only a few works [6,7,30–32] tackle with the problem of imbalanced classification utilizing deep learning in the literature. Most of them rely on shallow models and hand-crafted features. Khan et al. [31] proposed a cost-sensitive deep neural network to automatically learn robust feature representations for both the dominant and rare classes. To handle the problem of classifying imbalanced data, Jeatrakul et al. [30] proposed a method combined Synthetic Minority Over-sampling Technique (SMOTE) and Complementary Neural Network (CMTNN). To learn discriminative representation for imbalanced classification, Huang et al. [7] proposed a deep learning framework through quintuplet instance sampling and the associated triple-header hinge loss. Yan et al. [32] proposed a learning framework to improve multimedia data classification, in which CNNs are integrated with a bootstrapping sampling algorithm which creates a set of balanced training batches, each with a few positive instances. Most of these methods can be classified as natural extensions to traditional algorithms to handle imbalanced data classification. In this paper, we focus on tackling the problem of imbalanced data classification based on ensemble learning, combined with deep learning.

---

1 http://image-net.org/challenges/LSVRC/2014/
2 http://image-net.org/challenges/LSVRC/2015/ and http://mscoco.org/dataset/#detections-challenge2015

 doi:10.20944/preprints201709.0037.v1

1.3. Ensemble Learning

Ensemble Learning is a hot topic in machine learning. In machine learning, an ensemble contains a number of learners which are commonly called base learners. Base learners are also referred as "week learners", for they are usually slightly better than random guess. Ensemble learning utilize a set of learning algorithms to obtain better classification results than could be obtained from any of the constituent learning algorithms alone. During the train period, multiple classifiers are employed to learn the original dataset respectively in ensemble classification learning to solve the same problem. The results from the trained ensemble will be combined and then used to classify the unknown data. The functions of single model have high classification performance but have a problem in terms of a fixed a set of parameters, which causes the bias. Reduction of such bias can be obtained through the ensemble learning. A comprehensive review of ensemble learning can be found in [33].

As mentioned above, the generalization ability of an ensemble is usually much stronger than that of base learners. Ensemble learning is able to boost week learners to strong learners which can make very accurate predictions. Thus, ensemble learning is appealing in machine learning. An ensemble is a supervised learning algorithm essentially, for it can be trained and then used to make prediction for unknown data. The performance of ensemble learning depends on the precision of the constituent classifiers. Although most theoretical analyses work on weak learners, it is notable that base learners used in practice are not necessarily weak since using not-so-weak base learners often results in better prediction performance.

Ensemble Learning can be mainly categorized into three types as follows:

- Bagging

  Bagging is the abbreviation of "bootstrap aggregating", which was proposed by Leo Breiman[34] to improve the classification by combining prediction results of models trained independently on randomly generated training sets. That is to say, bagging is a special case of the model averaging approach, involves having each model in the ensemble vote with equal weight.. In order to improve the stability and accuracy, bagging usually trains each model in the ensemble using a randomly drawn subset of the training set. Moreover, the random selection of training set can also reduces variance and helps to avoid overfitting. As an example, the random forest [35] algorithm combines a collection of random decision trees with bagging to achieve high classification accuracy.
- Boosting

  Boosting is a machine learning ensemble meta-algorithm, based on the question: *Can a set of weak learners create a single strong learner* [36]. It involves incrementally building an ensemble by iteratively training a new model instance to emphasize the training samples which previous models misclassified. Most boosting algorithms consist of iteratively learning weak classifiers with respect to a sampling distribution of the training instances, and adding them to form a final strong classifier. When they are added, they are typically weighted in some way that is usually related to the weak learners' accuracy. To stress the training samples which previous models misclassified, after a weak learner is added, training instances that are misclassified gain weight and training instances that are classified correctly lose weight. Then training instances that previous models misclassified are more prone to be drawn during the period of training for the next weak learner to be added into the ensemble. Although many newer algorithms are reported to yield better results, the most usual implementation of boosting is still Adaboost [37] by far.
- Bucket of models

  A "bucket of models" is a ensemble learning technique in which a model selection algorithm is utilized to choose the best model in a set for each problem. On average, when evaluated across a large of problems, it will typically yield much better results than any model in the bucket. The most common implementation of model selection is Cross-Validation Selection. In order to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2017 doi:10.20944/preprints201709.0037.v1

Figure 2. The proposed image classification framework for images acquired from visual traffic surveillance senso.

choose the best model, Cross-Validation Selection try all models int the set with the training data, and pick the one that has the best performance for problems. However, when a set of models is used with a large set of problems, it may be desirable to avoid training some of the models that take a long time to train. To avoid the time consuming training, a meta-learning approach named Landmark learning [38] is often utilized. It involves training only the fast algorithms in the bucket, and then using the performance of these imprecise algorithms to help determine which slow algorithm is most likely to do best for a large set of problems.

2. The proposed scheme

In this section, we presented the proposed deep learning framework for vehicle type classification on visual traffic surveillance sensors and the whole framework is showed in Figure 2. First, a balanced sampling data augmentation strategy is used to increase the number of samples of rare classes in the original dataset, which can reduce classification bias and use as much data as possible for training. Then, a set of convolutional neural networks models are trained on the balanced data set, all started from a good initialization *(pretrained on ImageNet)*. Finally, outputs of multiple models are combined together by maximum voting policy according to the predictions of single models. The details of the framework is presented as follows.

Input:
an original imbalanced data set $\mathcal{D}$;
rare classes $\mathcal{C} = \{c_1, c_2, \cdots, c_m\}$;
train data of rare classes $\mathcal{D}_r = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_m\}$;
train data of rare classes after sampling $\mathcal{D}_r^{'} = \emptyset$;
an threshold n;

Output:

an training data set after balanced sampling $\mathcal{D}^{'}$;
for $i = 0$ to m **do**
  $\mathcal{D}_i^{'} = \emptyset$
  $s = size(\mathcal{D}_i)$
  $\boldsymbol{s}$ = randperm(n)
  for $j = 0$ to n **do**
    $ind = \boldsymbol{s}[j] \bmod s$
    $\mathcal{D}_i^{'} = Concat(\mathcal{D}_i^{'}, \mathcal{D}_i(ind))$
  end for
  $\mathcal{D}_r^{'} = Concat(\mathcal{D}_r^{'}, \mathcal{D}_r^{'})$
end for
$\mathcal{D}^{'} = (\mathcal{D} - \mathcal{D}_r) \bigcup \mathcal{D}_r^{'}$

Algorithm 1: Balanced Sampling

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2017 doi:10.20944/preprints201709.0037.v1

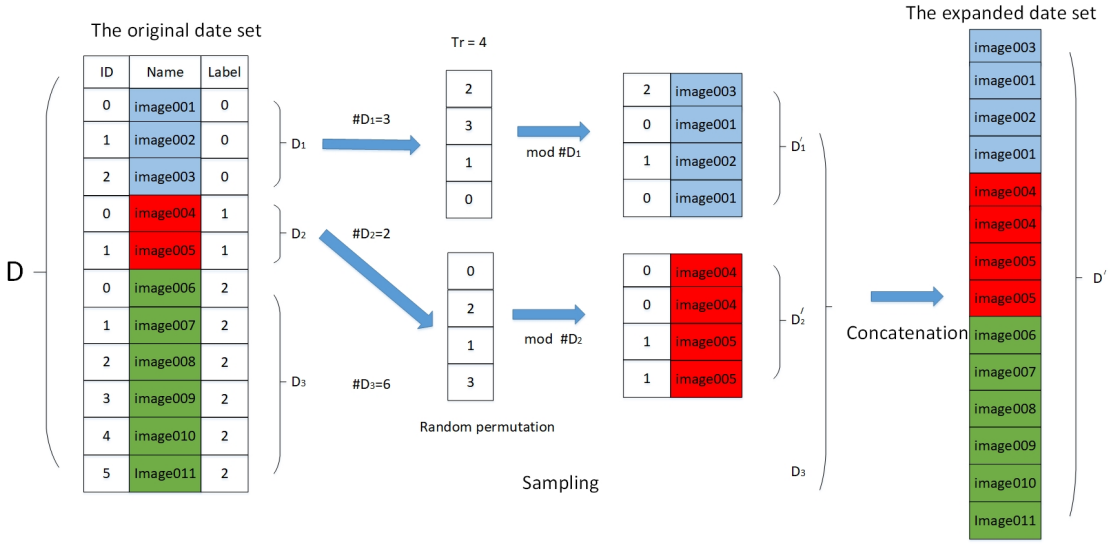

Figure 3. Balanced sampling

2.1. Data augmentation with balanced sampling

To ease problems caused by extreme imbalanced data distribution, we use over sampling with random shuffling. That is, the size of the minority class is increased randomly by over-sampling. Nevertheless, since this method replicates existing examples in the minority class, over-fitting is more likely to occur. To avoid over-fitting, the size of the minority class is increased to a small number, compared with the size of the majority class in practical application. The details of the proposed balanced sampling scheme is shown in Fig.3, let $Tr = n$ be a threshold that denotes the size each rare class will be increased to. For the rare class c_i, firstly a random permutation $S = (S[0], S[1], \cdots, S[n])$is generated. Then, we get the actual identity for $S[j]$ with the following equation:

$$ind = S[j] \, mod \, n, \tag{1}$$

where *mod* denotes the modulo operation. After selecting samples by a random permutation and modulo operations, we get the expanded $\mathcal{D}_i^{'}$ for the rare class c_i based on the original data set $\mathcal{D}_i$. At last, the samples of all rare classes and the other classes are concatenated and reshuffled. The details of the proposed balanced sampling scheme is presented in Algorithm 1.

2.2. Revisiting ResNets

In this subsection, we briefly introduce the ResNets used in this paper. To ease the training of framework that are substantially deeper than those employed previously, He et al. [29] proposed a residual learning framework named ResNets. Deep residual networks consist of many stacked Residual Units as shown in Figure 4. Each unit can be expressed in a general form [39]:

$$\begin{aligned} y_l &= h(x_l) + F(x_l, W_l), \\ x_{l+1} &= f(y_l), \end{aligned} \tag{2}$$

where x_l and x_{l+1} are input and output of the l-th unit, and F is a residual function. In [29], f is a ReLU [40] function, and $h(x_l) = x_l$ is an identity mapping.

For ResNets with units in Figure 4(b) is much easier to train and has a better generalization than the original ResNets in [29], we use ResNets in [39] in this paper.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2017 doi:10.20944/preprints201709.0037.v1

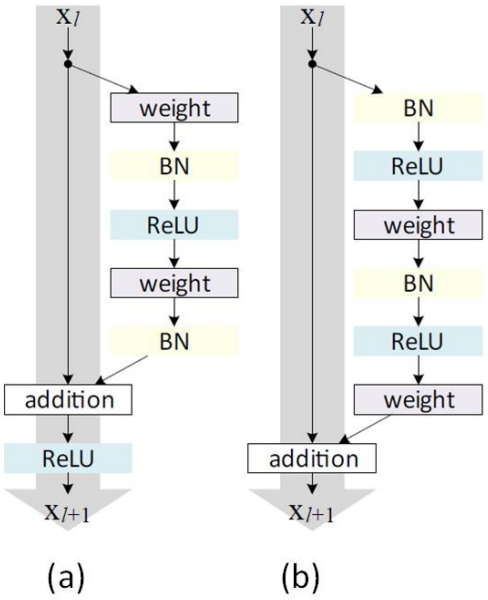

Figure 4. (a) Residual Unit in [29]; (b) Residual Unit in [39].

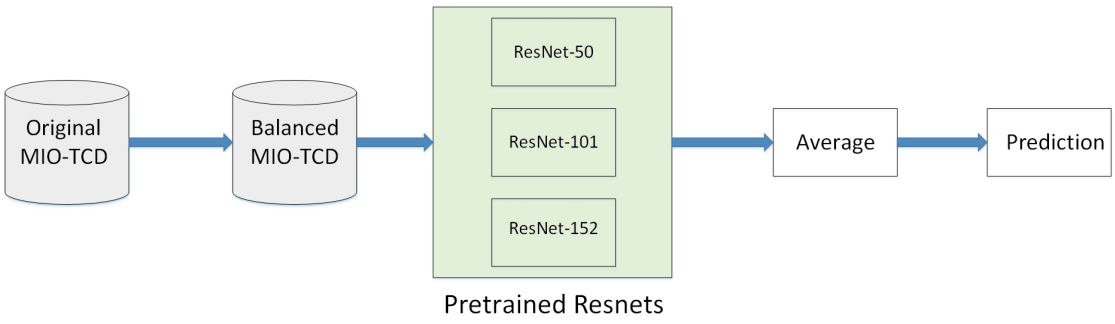

Figure 5. The framework of our deep CNN ensemble model.

2.3. Deep ensemble model

To tackle the imbalanced classification problem for Vehicle Type Classification, we proposed an deep CNN ensemble model. The deep CNN ensemble contains ResNet-50, ResNet-101 and ResNet-152. As is shown in 5, the proposed ensemble model comprises three key stages: starting CNN models from good initial parameters, fine tuning network parameters and averaging models. Concretely, firstly all of CNN models in the ensemble are pretrained on ImageNet. Next, the network parameters are refined using MIO-TCD data set enhanced by data augmentation. Finally, the outputs of refined CNN models are combined together by averaging their predictions.

As mentioned above, the proposed ensemble model contains multiple deep learning models. Therefore, the initial stage of the ensemble system generates many results for single a image to classify. The voting process is necessary to decide which class the image belongs to based on votes. In this paper, maximum majority voting is adopted to classify images based on initial predictions of single models. For the number of models in our ensemble is odd, consequently it doesn't have to be considered for cases with same votes in this paper.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2017 doi:10.20944/preprints201709.0037.v1

Table 1. Number of training samples for each category in MIO-TCD Dataset

Category	#
Articulated truck	10,346
Background	160,000
Bicycle	2,284
Bus	10,316
car	260,518
Motorcycle	1,982
Non-motorized vehicle	1,751
Pedestrian	6,262
Pickup truck	50,906
Single unit truck	5,120
Work van	9,679
Total	519,164

3. Experiments and results

3.1. Details of the MIO-TCD classification challenge dataset

To demonstrate the effectiveness of our proposed framework, we use the MIO-TCD classification challenge dataset[3] for testing, which is a large benchmark traffic camera data set with a highly imbalanced data distribution. The dataset consists 648,959 samples in the classification dataset acquired at different times of the day and different periods of the year by traffic cameras deployed all over Canada and the United States. Those images have been selected to cover a wide range of challenges and are representative of typical visual data captured in urban traffic scenarios.

The classification challenge dataset contains 648,959 images divided into 11 categories, including *Articulated truck*, *Background*, *Bicycle*, *Bus*, *Car*, *Motorcycle*, *Non-motorized vehicle*, *Pedestrian*, *Pickup truck*, *Non-motorized vehicle*, *Single unit truck* and *Work van*. The size of training samples is 519,164. The number of training samples for each category is given in Table 1. As is shown in Table 1, number of samples for each category in MIO-TCD Dataset is in a range between 1,751 and 260,518. *Bicycle*, *Motorcycle* and *Vehicle* categories only contain a small number of training samples, while *Background* and *Car* make up the majority.

3.2. Evaluation criterion

The prime goal of this paper is to introduce a new vehicle type classification scheme on the images acquired from multi-view Visual Traffic Surveillance Sensors. In order to objectively evaluating the performance of the introduced method, we evaluate our approach by the following 6 metrics.

- **Precision of each category**

$$Pre_i = \frac{TP_i}{TP_i + FP_i}$$

- **Recall of each category**

$$Rec_i = \frac{TP_i}{TP_i + NF_i}$$

- **Accuracy**

$$Acc = \frac{TP}{\#ofTestingImages}$$

- **Mean Recall**

$$mRe = mean(Rec_i)$$

[3] http://tcd.miovision.com/challenge/dataset/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2017 doi:10.20944/preprints201709.0037.v1

Table 2. Comparisons of precision for each category on the MIO-TCD Dataset. AT denotes articulated Truck, MC denotes Motorcycle,NV denotes Non-motorized Vehicle, PT denotes Pickup Truck, SUT denotes Single Unit Truck, WV denotes Work Van, and BG denotes Background.

Model	AT	Bicycle	Bus	Car	MC	NV	Pedestrian	PT	SUT	WV	BG
ResNet-50	0.8748	0.9903	0.8135	0.9712	0.9718	0.8971	0.6516	0.9007	0.8644	0.7158	0.9013
ResNet-50-BS	0.8976	0.9930	0.8336	0.9754	0.9717	0.9180	0.5211	0.9048	0.8845	0.6814	0.9312
ResNet-101	0.8986	0.9923	0.8401	0.9772	0.9828	0.9320	0.7387	0.9313	0.8915	0.7450	0.9283
ResNet-101-BS	**0.9314**	0.9926	**0.8632**	0.9809	0.9806	0.9421	0.6466	0.9469	0.9089	0.7271	0.9315
ResNet-152	0.9050	0.9935	0.8471	0.9781	**0.9835**	0.9100	0.7390	0.9336	0.8870	0.7624	0.9368
ResNet-152-BS	0.9146	**0.9939**	0.8424	**0.9850**	0.9813	0.9287	0.6435	0.9421	0.9118	0.7403	0.9384
DCEM	0.8936	0.9923	0.8581	0.9765	0.9820	0.9439	**0.7812**	0.9392	0.9124	0.7844	0.9526
DCEM-BS	0.9115	0.9929	0.8269	0.9811	0.9830	**0.9550**	0.7538	**0.9718**	**0.9445**	**0.8328**	**0.9679**

Table 3. Comparisons of recall for each category on the MIO-TCD Dataset.

Model	AT	Bicycle	Bus	Car	MC	NV	Pedestrian	PT	SUT	WV	BG
ResNet-50	0.8829	0.9939	0.8021	0.9411	0.9722	0.8808	0.3288	0.8754	0.8973	0.7063	0.7878
ResNet-50-BS	0.8469	0.9924	0.8861	0.9383	0.9759	**0.9272**	0.5913	0.9048	0.8937	0.7805	0.7655
ResNet-101	0.9076	0.9961	0.8739	0.9636	0.9761	0.9131	0.4840	0.9093	0.9364	0.7852	0.8390
ResNet-101-BS	0.8713	0.9962	0.8949	0.9581	0.9799	0.9212	0.5890	**0.9233**	0.9281	**0.8266**	0.8369
ResNet-152	0.9026	0.9956	0.8827	0.9686	0.9755	0.9192	0.4977	0.9157	**0.9408**	0.7898	**0.8625**
ResNet-152-BS	0.8740	0.9955	0.8984	0.9647	0.9809	0.9212	**0.6142**	0.9157	0.9287	0.8219	0.8551
DCEM	0.9192	0.9968	0.8687	0.9655	0.9816	0.9172	0.4566	0.9176	0.9346	0.7758	0.8456
DCEM-BS	**0.9312**	**0.9984**	**0.9037**	**0.9663**	**0.9889**	0.9010	0.5594	0.9022	0.9402	0.7898	0.8468

- **Mean Precision**

$$mPre = mean(Pre_i)$$

- **Cohen Kappa Score**

$$k = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio), and p_e is the expected agreement when both annotators assign labels randomly [41].

3.3. Baselines

To indicate the effect of the proposed scheme, the state of art deep learning methods ResNet-50, ResNet-101 and ResNet-152 [29] are used. The ResNet-50 trained with balance sampling is denoted as ResNet-50-BS, by analogy to ResNet-101-BS and ResNet-152-BS. We name the proposed method with *DCEM-BS*. The comparison experiment results are presented as follows.

3.4. Results

Table 2 presents comparisons of precision for each category on the MIO-TCD classification challenge dataset. Table 2 indicates that the proposed method *DCEM-BS* got the best performance in term of precision for each category as a whole, from comparisons with the baselines. Moreover, networks with balanced sampling got better performances than the others. Table 2 shows that ensemble learning and balanced sampling is effective to improve precision of vehicle classification.

Table 3 presents comparisons of precision for each category on the MIO-TCD classification challenge dataset. It indicates that balanced sampling is able to improve the recall for vehicle classification obviously, particularly for classes that is not dominant such as *Pedestrian* and *Work van*

We got 0.8844 mean recall, 0.9776 classification accuracy, 0.9201 mean precision, and 0.9651 Cohen Kappa Score on verification data. The performance comparison with other deep learning models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2017 doi:10.20944/preprints201709.0037.v1

Table 4. The overall results on the MIO-TCD Dataset

Model	Mean Recall	Precision	Mean Precision	Cohen Kappa Score
ResNet-50	0.8244	0.9586	0.8684	0.9354
ResNet-50-BS	0.8639	0.9610	0.8648	0.9392
ResNet-101	0.8713	0.9691	0.8713	0.9520
ResNet-101-BS	0.8841	0.9705	0.8956	0.9540
ResNet-152	0.8773	0.9698	0.8978	0.9531
ResNet-152-BS	**0.8882**	0.9713	0.8929	0.9553
DCEM	0.8708	0.9723	0.9106	0.9568
DCEM-BS	0.8844	**0.9776**	**0.9201**	**0.9651**

are shown in 4, which demonstrate the proposed scheme is able to increase mean precision to some extend, compared with the baseline algorithms. Concretely, the proposed DCEM-BS improves the mean precision by more than 2% in contrast to the single models. Moreover, DCEM-BS is better than DCEM in terms of performance, which indicates that our balanced sampling tactic is effective. More results of the proposed scheme can be found in the classification challenge related web page. [4].

4. Conclusion

To correctly classify vehicle type on images acquired from visual traffic surveillance sensors, we proposed a image classification scheme based on ensemble deep learning. Experiments on the MIO-TCD classification challenge dataset demonstrate that the proposed method is able to increase mean precision to some extend, compared with the baseline algorithms.

5. Acknowledgement

This work is supported by the Nature Science Foundation of China(No.61662024, No.61602220), the Natural Science Foundation of Jiangxi Province (20171BAB212013, 20161BBI90004, 20161BAB212057), and the Open Fund of Hubei Province Key Laboratory(2016KLA03) .

References

1. Liu, W.; Ji, R.; Li, S. Towards 3D object detection with bimodal deep boltzmann machines over RGBD imagery. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3013–3021.
2. Liu, W.; Li, S.; Lin, X.; Wu, Y.; Ji, R. Spectral–spatial co-clustering of hyperspectral image data based on bipartite graph. *Multimedia Systems* **2016**, *22*, 355–366.
3. Liu, W.; Li, S.; Cao, D.; Su, S.; Ji, R. Detection based object labeling of 3D point cloud for indoor scenes. *Neurocomputing* **2016**, *174*, 1101–1106.
4. Zadrozny, B.; Langford, J.; Abe, N. Cost-sensitive learning by cost-proportionate example weighting. Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003, pp. 435–442.
5. Unsworth, C.; Coghill, G. Excessive noise injection training of neural networks for markerless tracking in obscured and segmented environments. *Neural computation* **2006**, *18*, 2122–2145.
6. Zhou, Z.H.; Liu, X.Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering* **2006**, *18*, 63–77.
7. Huang, C.; Li, Y.; Change Loy, C.; Tang, X. Learning deep representation for imbalanced classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5375–5384.
8. Chen, C.; Liaw, A.; Breiman, L. Using random forest to learn imbalanced data. *University of California, Berkeley* **2004**, *110*.

[4] http://podoce.dinf.usherbrooke.ca/methods/classification/199/

9. Tang, Y.; Zhang, Y.Q.; Chawla, N.V.; Krasser, S. SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **2009**, *39*, 281–288.
10. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **2002**, *16*, 321–357.
11. Drummond, C.; Holte, R.C.; others. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Workshop on learning from imbalanced datasets II. Citeseer Washington DC, 2003, Vol. 11.
12. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Advances in intelligent computing* **2005**, pp. 878–887.
13. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* **2009**, *21*, 1263–1284.
14. Maciejewski, T.; Stefanowski, J. Local neighbourhood extension of SMOTE for mining imbalanced data. Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on. IEEE, 2011, pp. 104–111.
15. MetaCost, D. A General Method for Making Classifiers Cost—sens˜ ive. Proc. of the 5th International Conference on Knowledge Discovery and Data Mining. San Diego, USA:[sn], l999.
16. Ting, K.M. A comparative study of cost-sensitive boosting algorithms. In Proceedings of the 17th International Conference on Machine Learning. Citeseer, 2000.
17. Yan, Y.; Liu, Y.; Shyu, M.L.; Chen, M. Utilizing concept correlations for effective imbalanced data classification. Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on. IEEE, 2014, pp. 561–568.
18. Batista, G.E.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter* **2004**, *6*, 20–29.
19. Zhang, L.; Wang, W. A re-sampling method for class imbalance learning with credit data. Information Technology, Computer Engineering and Management Sciences (ICM), 2011 International Conference on. IEEE, 2011, Vol. 1, pp. 393–397.
20. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **2009**, *39*, 539–550.
21. Sun, A.; Lim, E.P.; Liu, Y. On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems* **2009**, *48*, 191–201.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, pp. 1097–1105.
23. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
24. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
25. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
26. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv preprint arXiv:1312.4400* **2013**.
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
28. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016, pp. 770–778.
30. Jeatrakul, P.; Wong, K.; Fung, C. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. *Neural Information Processing. Models and Applications* **2010**, pp. 152–159.
31. Khan, S.H.; Bennamoun, M.; Sohel, F.; Togneri, R. Cost sensitive learning of deep feature representations from imbalanced data. *arXiv preprint arXiv:1508.03422* **2015**.
32. Yan, Y.; Chen, M.; Shyu, M.L.; Chen, S.C. Deep learning for imbalanced multimedia data classification. 2015 IEEE International Symposium on Multimedia (ISM). IEEE, 2015, pp. 483–488.
33. Zhou, Z.H. Ensemble learning. *Encyclopedia of biometrics* **2015**, pp. 411–416.

 doi:10.20944/preprints201709.0037.v1

372 34. Breiman, L. Bagging predictors. Univ. California Technical Report No. 421 **1994**.
373 35. Ho, T.K. Random decision forests. Document Analysis and Recognition, 1995., Proceedings of the Third
374 International Conference on. IEEE, 1995, Vol. 1, pp. 278–282.
375 36. Kearns, M.; Valiant, L. Cryptographic limitations on learning Boolean formulae and finite automata.
376 *Journal of the ACM (JACM)* **1994**, *41*, 67–95.
377 37. Wang, L.; Sugiyama, M.; Yang, C.; Zhou, Z.H.; Feng, J. On the Margin Explanation of Boosting Algorithms.
378 COLT, 2008, pp. 479–490.
379 38. Barto, A.G.; Sutton, R.S. Landmark learning: An illustration of associative search. *Biological Cybernetics*
380 **1981**, *42*, 1–8.
381 39. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. European Conference on
382 Computer Vision. Springer, 2016, pp. 630–645.
383 40. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. Proceedings of the
384 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.
385 41. Artstein, R.; Poesio, M. Inter-coder agreement for computational linguistics. *Computational Linguistics* **2008**,
386 *34*, 555–596.