Optimization of Regional - Convolutional Neural Network for Electricity Conservation Using Arduino

S. Ushasukhanya  
Department of Computer Science and Engineering, Annamalai University, India.  
E-mail: ushasukhanya@gmail.com

S. Jothilakshmi  
Department of Information Technology, Annamalai University, India.  
E-mail: jothi.sekar@gmail.com

Received October 06, 2020; Accepted November 15, 2020  
ISSN: 1735-188X  
DOI: 10.14704/WEB/V18SI01/WEB18005

Abstract

The demand for electrical energy in developing countries is apparently increasing thereby creating a large gap between the availability of the electrical resource and its growing demand. Globally reputed energy economists have recognized that 25% of reduction in energy consumption can be achieved by adopting efficient energy conservation techniques. This paper presents one of the simplest ways of conservation techniques that enables the electric power supply only when it is actually needed. It is an automatic system that functions with the existing CCTV surveillance camera to enable/disable the electric power supply, only in the location where human is present / absent respectively. The proposed approach is demonstrated without the use of sensors, based on Regional Convolutional Neural Network (R-CNN). A new R-CNN model is constructed for CHOKEPOINT dataset and the optimization is done using Nadam technique. The results are then fed into Arduino micro controller to control the electric supply based on the presence/absence of human in the particular region.

Keywords

Smart Electricity Conservation Technique, R-CNN, Nadam, Arduino.

Introduction

Electricity is the highly accomplished and most composed form of energy. Its usage is eco-friendly as it is a non-polluting resource and hence its demand is highly increasing. According to Verified Market Research, the Global Power Electronics Market was valued
at USD 36.35 Billion in 2017 and is projected to reach USD 55.07 Billion by 2025, growing at a compound annual growth rate (CAGR) of 5.4% from 2018 to 2025. The current construction standards yield a very poor performance in electricity conservation [1] and hence many devices are now being developed by the power electronics branch to conserve electricity. In spite of many measures to save electricity is being carried out, a technique to meet the demand is still one of the most significant challenges [2-6]. Figure 1 shows the energy consumption trend for the indoor environment in India which is likely to increase faster than other countries. Most of the conservation methods use sensors to control the electric power supply [7], which places an additional installation cost. Hence, this paper aims at developing one of the simplest techniques to conserve electricity in an indoor environment by switching on the electric supply only when human is detected in CCTV footage. It is also location specific, where the electricity even for a part of the room is controlled.

![Figure 1](image1.png)

**Figure 1 Energy consumption trend**

The entire system is divided into two phases. Effective human detection in CCTV footage which is succeeded by control of electric power supply using Arduino micro-controller. The existing Closed circuit television (CCTV) video footage is converted at the rate of 21 frames per second. The frames that has a motion or movement detected are extracted and pre-processed to enhance the quality of the image. Data augmentation techniques like skew tilt and random erase are carried out to improve the classification accuracy. A R-CNN model is constructed to classify these images as images with and without human along with the bounding box specifying the location of the human being. Dropout is the regularization technique used to generalize the model with the final optimization done by
Nadam technique. These results are then fed into the Arduino micro controller which is connected to the electric power supply of the room to control the power supply based on human’s presence/absence.

The rest of the paper is structured as follows: Section 2 briefly describes the literature survey for the proposed work. The overview of the proposed framework is presented in Section 3. Each module of the proposed work is then described in detail. Experiments and results are discussed in Section 4, succeeded by conclusions and future enhancements in Section 5.

Related Work

An automated approach is proposed in this work for electricity conservation by disabling the electric power when it is actually not needed. Hence, techniques for identification and localization of human in CCTV images using deep learning techniques are addressed in this section. In addition, data augmentation and optimization techniques for enhancing the training process and improving the classification accuracy are also discussed here with their benefits and limitations.

Deep Learning Techniques for Human Identification

Deep learning techniques are now being widely used in various applications for designing automatic systems. Along with detecting the object, the region of the object being present is also predicted using region-based classifiers like R-CNN. Figure 2 shows the example architecture of R-CNN designed for this approach. Classifiers like R-CNN have now been used for various predictions in the agriculture field from rainfall and soil fertility prediction to harvesting techniques.

![Figure 2 Classification of images by R-CNN](http://www.webology.org)
One of the techniques of apple harvesting is the shaking of apple tree branches and catching the apples that fall from the tree on shaking. This is done to overcome the labour cost, and one such prediction is designed by Jing Zhang [8]. The system uses Pseudo-Color Image and Depth (PCI-D) images with Regional Convolutional neural network (R-CNN) to identify and localise the apple tree branches. The input for R-CNN is the features of apple tree branches. The pseudo-colour images, depth images and RGB images in a real habitat are obtained by Microsoft Kinect v2.

R-CNN then uses selective search (SS) algorithm to compute region proposals, which is followed by CNN for classification and bounding-box regression. The average accuracy of the PCI method is 81%, which is lesser than that of PCI-D, which yields 86%. The correlation coefficient values of PCI and PCI-D is 0.91 and 0.86, respectively. These values are obtained from the results of branch detection, and hence it is proved that the combination of pseudo-colour images with depth images yields good results in identifying branches, locating branches and performing skeleton fitting on them. Future enhancement suggested in this paper is to incorporate modified Faster R-CNN method to improve the detection rate of apple tree branches.

As the R-CNN is considered to be slower than the other single stage regional based classifiers, a new two-stage detector based on R-CNN was developed by Zeming li et al [9]. In this system, the head of network is made as light as possible, by using a large-kernel separable convolution to produce a thin feature map and an R-CNN subnet. The system greatly reduces the computation time thereby reducing the memory consumption. Light Head R-CNN gets 30.7 mAP at 102 FPS on COCO, there by exceeding the performance of other regional based models like YOLO and SSD. This is achieved by just embedding the model with a tiny network [10] [11]. Though object detection from CCTV images is being performed for the last few decades, there are still existing challenges like low resolutions, strong illuminations, pose variations etc. Chen chen Zhu et al [12] designed a system to overcome the above said challenges using a contextual multi scale-RCNN for face detection. The network considers the multi-scale features for its prediction. In addition to this, face representation is done by synchronizing the location specific features and semantic features of low-level and high-level layers respectively. The experimental results on face detection Data Set and Benchmark (FDDB) proves that it outperforms other baseline methods by a higher margin.
Optimization of Deep Neural Networks

Although CNN and its variants yields a good performance in classification, it requires high computational resources. Hence the need for optimizing the network has become a more popular task. Various researchers are working in optimizing the neural networks and one such system to solve these optimization problems, is developed by Sun et al [13]. A new system called particle swarm optimization with binary encoding (BQPSO), developed by adjusting the evolution to discrete binary space, is created in the following equations. The movement of the particles $M$ depends on the following equations:

$$m_{best} = \left( \frac{1}{M} \sum_{i=1}^{M} P_i(t), \frac{1}{M} \sum_{i=1}^{M} P_i(t), ..., \frac{1}{M} \sum_{i=1}^{M} P_i(t) \right)$$  \hspace{1cm} (1)

$$a_{id}(t) = \varphi \times p_{id} + (1 - \varphi) \times p$$  \hspace{1cm} (2)

$$x_{id}(t + 1) = a_{id} \pm \beta |m_{best, id} - x_{id}(t)|^{\alpha} \ln \left( \frac{1}{u} \right)$$  \hspace{1cm} (3)

Where $P$ is the particle, $\varphi$ and $u$ are random numbers distributed uniformly on $[0,1]$, $\beta$ is a positive constant called Contraction Expansion Coefficient. Inspired by this Li Yangyang et al [14] optimized the CNN using Particle swarm optimization method (PSO) with Genetic algorithm (GA). He has integrated binary encoding with PSO thereby incorporating the strong features of genetic algorithm. In addition to it, a new measurement method is used to bridge the gap between classification accuracy and computational cost. Ultimately this technique has increased the speed of the computations at no additional cost.

Georgia et al [15] proposed a new Mesh-RCNN system to detect objects in real world using images and to give a full 3D structure of the detected image. The 2D recognition and augmentation is done by building the Mesh-RCNN on mask R-CNN. It produces high resolution triangle meshes which should effectively modify their complexity, topology, and geometry in response to varying visual stimuli. Similar works for 2D recognition have been done in [16][17][18]. The three staged refinement branch has six convolution layers organized into three residual blocks. Adam optimization [19] is used for fine-tuning the parameters of the model for better efficiency. 25 epochs of training with learning rate $10^{-4}$ and 32 images per batch is used on 8 Tesla V100 GPUs. The experiment has been validated on Shape Net and the merits are shown in PiX3D.

Deep neural networks optimization has also gained importance in medical field for various applications and one such application is liver segmentation. Liver transplantation requires enormous pre-evaluation strategies for which computer-aided-diagnosis is
mandatory. The images from various modalities like Computed tomography (CT) scanning and magnetic resonance imaging (MRI) differ from each other due to various difference in settings. Hence a system to segment, identify and recognize certain things from these images are essential. One such system using deep neural network was designed by Ram and Mohana sankar et al [20] for liver segmentation. The system combines holistically-nested edge detection (HED) and Mask-region-convolutional neural network (Mask R-CNN). To avoid the over fitting issues in Mask R-CNN, Adam optimizer is used which also boosts the training process. The multi-task loss function of Mask R-CNN combines the loss of classification, localization and segmentation mask.

\[
Total \ Loss = Loss_{cls} + Loss_{box} + Loss_{mask} \tag{4}
\]

Where \(Loss_{cls}\) is classification loss, \(Loss_{box}\) is bounding box regression loss and \(Loss_{mask}\) is mask loss.

Experiments

Data Set

CHOKEPOINT is a video dataset designed for experiments in person identification/verification under real-world surveillance conditions. It consists of 64,204 images of 48 video sequences in three different classes of indoor surveillance images in which 25 and 29 subjects of two different portals are considered. These subjects in two different portals are recorded with one month apart with the frame rate of 30 per second, and the resolution of 800×600. Nearly, 70% of the images are considered for training set and the remaining is considered for test set.

Data Augmentation

Two main issues that hinder the performance of the neural network in computer vision is image inconsistency and number of training images. To overcome this, there are many data augmentation techniques that increases the overall prediction of the model. Single augmentation technique like Skew tilt and Random erase seem to perform well and hence they are used in this work. All the training images are tilted forwards, backwards, left, or right a maximum of 23.5°. This gives the illusion that the image is being viewed from a different perspective than originally seen and creates realistic examples. This is followed by Random erase in which the pixels of the image within a random rectangular area are erased with random RGB values and the maximum area affected by the augmentation is 50% of the pixels. Label Image is the image annotation tool which is then used after
augmentation to annotate all the images with bounding-boxes and class labels. As a result, an XML file is generated with details about the annotated objects and bounding boxes. These results are used by the model for further processing.

**Pre-processing**

The background subtraction, is the method of removing the background image from the real image for the purpose of easy video sequences processing. The regions of interest (ROIs) is found by applying the Gaussian mixture model (GMM) for background subtraction of video frames [21]. Then the noise is eliminated using certain morphological filtering, such as closing and opening [22], which represents the required object areas in rectangles. The resultant image after background subtraction is given by equation 9.

\[ O(x, y) = I(x, y) - B(x, y) \]  

**R-CNN**

The proposed work requires not only detection of human but also the location where the human is actually present to make the electricity turn on/off, particularly in that location. As Regional-Convolutional Neural Network (R-CNN) does object detection along with identifying the location of the object, which is represented by bounding boxes containing the objects, it is used in this experiment. R-CNN uses selective search (SS) algorithm for generating region proposals [23]. For every image around 2k region proposals are generated which is done by the bottom-up merging of regions. CNN is then used to compute a feature vector, which generates about 4096-dimensional vector for each proposal [24]. The last step is the classification of images, which is done by Support Vector Machine.

**Selective Search**

Selective search is used to obtain all possible set of object locations in image processing techniques. It is preferred for its high recall value. One of the popular approaches for segmentation is bottom-up merging [25]. The over-segmentation method is used as the root of this process. All bounding boxes corresponding to the segmented parts of the region are summed up using a greedy algorithm. The adjacent segments are merged based on the similarity score of colour, shape, texture and size. Colour similarity is calculated by

\[ S_{colour}(X_i,X_j) = \sum_{n=1}^{m} \min(a_i^n, a_j^n) \]
Where $X_i$ and $X_j$ are the two regions, $a_i^m$ is the histogram value for $m^{th}$ bin in colour descriptor. Texture similarity is computed by

$$S_{texture}(X_i, X_j) = \sum_{m=1}^{n} \min\left(b_i^m, b_j^m\right)$$  \hspace{1cm} (7)

Where the histogram value for $m^{th}$ bin is, $b_i^m$. Size similarity is calculated by

$$S_{size}(X_i, X_j) = \frac{1-size(x_j) + size(x_j)}{size(i)}$$  \hspace{1cm} (8)

Where the size of image pixels is represented by $size(i)$. The shape similarity is measured by

$$S_{shape}(X_i, X_j) = \frac{1-size(GG_{ij}) - size(x_i) + size(x_j)}{size(i)}$$  \hspace{1cm} (9)

Where size ($GG_{ij}$) is the bounding box around $x_i$ and $x_j$. The overall similarity is the addition of all these similarity scores. The process is repeated by calculating the similarity score between the resulting region and its neighbours for all regions until one single region is obtained for the entire image. Initially, all the locations are ranked, as there are better chances for duplicate boxes to obtain a high rank. The lower ranked boxes are filtered and if many merging methods recommend the same location, its possibility to arise from a visually coherent part of the image is very high [26] and hence, it is acceptable. Subsequently, 2000 candidate region proposals generated by selective search algorithm are encased and passed into CNN. The CNN acts as a feature extractor and generates 4096-dimensional feature vectors for each image. Finally, rather than trying to classify many regions, it just classified 2000 regions.

As it is strenuous to match these regions proposals with the labelled frames, Intersection over Union (IoU) is used to figure out accuracy of these regional proposals. IoU is defined as the ratio between the Predicted Box (PB) and the Ground-truth Box (GB), as shown in (10).

$$IOU = \frac{PB \cap GB}{PB \cup GB}$$  \hspace{1cm} (10)

Based on few trials, the threshold for this parameter is set to be 0.4 and thus when the IoU value is or exceeds 0.4, the region proposal contains the object, else it is classified as background.
Detection and Classification

The transfer learning models yields good performance only when trained with datasets similar to Image Net, and hence a new R-CNN model is developed. The first step is the training process of R-CNN model with the raw and augmented images of CHOKEPOINT dataset. The generation of candidate regions for every input image is done by the input layer.

The input is evenly divided into two segments by the convolution layers. The convolution operation is done by finding the dot product between the input feature ($Z_i$) and output feature ($Z_O$). Rectified Linear Unit (ReLU) has gained more popularity [27] among CNN models and hence it is used as the activation function in this model. The expression of ReLU is $f(x) = \max(0,x)$ which is shown in

![Figure 3 Relu function](image)

The output of every layer is given by the formula

$$O = \frac{(W-K+P)}{S+1} \quad (11)$$

Where $O$ is the output of every layer, $W$ is the input size, $K$ is the kernel size, $P$ is the padding rate, and $S$ is the stride at which the operations are done. A 2x2 max pooling is used to aggregate information from the feature map and padding layers. It is then proceeded forward through five convolutional layers and finally enters into the fully
connected (FC) layers which produces the feature vector. In addition to this, four offset values to improve the bounding box’s precision is also predicted. These offset values are useful in adjusting the bounding boxes. The Soft max function layer gives the confidence score for the bounding boxes with human and the probability of each label is generated. SVM then uses these extracted features to classify the features into any one of the classes. Finally, the model classifies the presence or absence of human within a frame. After the model is trained, the test data images are fed into the model and the results are evaluated and optimized to increase the efficiency of the model.

**Optimization**

The model is composed of various parameters and hyper-parameters which has a major impact on the overall efficiency. Both these parameters can be tuned and the optimal set of values are obtained after several trial and error methods. This actually increases the training time of the model drastically. There always exists a trade-off between the model’s training time and the efficiency of the model. The following table consists of the hyper-parameter values obtained after much iteration. Finally, Nadam optimization technique is used with the learning.

Table 1 Parameter values of R-CNN

<table>
<thead>
<tr>
<th>S.No</th>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No.of Layers</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Filter Size</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>No.of Filters</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>Mini batch</td>
<td>256</td>
</tr>
</tbody>
</table>

Fixed to 0.0001. It is also noted that the training process is inversely proportional to the filter size beyond 5 and hence the filter size is restricted to 5. A dropout of 20% is the regularization technique used by setting the value to 0.2 to reduce the difference between training and testing accuracy results or to smoothen the curve. The model was trained for 50 epochs without changing the global learning rate. Recall and Precision are the two parameters used to evaluate the training results of each image. The recall is defined as the ratio of ground-truth results detected over the total ground-truth results where as precision is the percentage of correctly classified results. Accuracy is an important parameter for evaluating the results of object detection which is shown in table 2.

Table 2 Range Analysis

<table>
<thead>
<tr>
<th>S.No</th>
<th>Optimization</th>
<th>Nadam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Training-accuracy</td>
<td>0.87</td>
</tr>
<tr>
<td>2</td>
<td>Test-accuracy</td>
<td>0.86</td>
</tr>
<tr>
<td>3</td>
<td>Training-loss</td>
<td>0.41</td>
</tr>
<tr>
<td>4</td>
<td>Test-loss</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Management of Electric Power Supply

The proposed work uses an Arduino Uno microcontroller to control or manage the main switch of the electric power supply based on human's presence/absence. Designing and prototyping have become easier and inexpensive with the usage of Arduino, as Arduino uses a simplified version of C/C++ for programming [25] [26]. The Arduino MEGA 2560 is selected for this work because it has more digital and analog pins that can enable the control of more loads. The output of the classifier is connected to the Arduino Uno microcontroller. On detecting the humans in a frame by the classifier, the Arduino controller turns on the light and fan in that room. When the classifier does not detect the human’s presence or when the person leaves the room, it classifies the image as an image without human and the controller turns off the electric power supply. Hence the power supply in the room is managed, only when it is actually required. This is done sequentially for all the frames of the video footage. The entire coding part of this Arduino microcontroller is done in embedded C program.

Results

All the frames of the CHOKEPOINT dataset are classified into frames with and without human by R-CNN. Figure 3 shows the classification of frames along with localization of the detected human, indicated with the bounding box.

![Figure 3(a & b) Human detection in a room and in corridor](image1)

![Figure 3(c) Human not being detected](image2)

Figure 4(a) shows the accuracy percentage obtained for both training set and testing set with respect to epochs. 4(b) shows the accuracy obtained with respect to various dropout levels. Highest accuracy is obtained with dropout set to 0.2.
Finally, electricity in and around 6 metres of the human being detected is enabled by the Arduino micro-controller while disabling the power in rest of the area or places where human is undetected. Figure 5 (a) represents the electricity being turned on only in the region of human’s presence where as 5(b) shows the electricity being disabled on human’s absence.

Conclusion

Deep learning has been showing promising results in computer vision domain. Hence the main intention of this work is to develop an automated system that manages the electric power supply when required using deep learning techniques. The system uses CHOKEPOINT dataset which consists of CCTV frames in three different sessions. These frames are augmented to increase the performance of the classifier in the later stage. The frames are classified as frames with and without human using R-CNN. R-CNN, in
addition to classification, also indicates the region of human’s presence. The model is trained on CHOKESPACE frames and optimized with Nadam algorithm. Arduino microcontroller is finally used to control the electricity when a human is detected in a frame. This enables the electric power supply to be turned on only in the region where a human is detected. Therefore, the fans and lights within a distance of 6 metres around the person’s detected region are enabled and the rest are disabled automatically. The entire system uses the existing CCTV cameras installed, thereby avoiding the additional cost of sensors. The system is developed for CCTV images obtained in day time and hence processing of night time images and improving the model’s performance at less training time using hyper parameter optimization are considered as future work.

References


