DETECTING PEOPLE IN ARTWORK BY CNNs

Introduction

With the increasing demand and advancement in CNNs, the challenges to CNNs are also increasing. CNNs are now used to solve problems which seemed almost impossible in the past. CNNs are widely used in detection of image recognition using datasets of images. This research paper deals with a more complex and interesting dataset, tackling a very challenging problem. The dataset contains images of different artwork and styles unlike just the common style like photos. This dataset is used for object detection of cross-depiction problem where the objects which are depicted in various forms are selected.

Dataset

A dataset called ‘People-Art’ is collected which consists of images, photos, cartoons from different artworks with only one class which is human. Humans are more commonly found in different artworks compared to other classes and henceforth this class is chosen. The dataset has 43 different styles out of which 41 are from WikiArt.org. The cartoons are collected from google searches, and the images are from PASCAL VOC 2021. The dataset is complicated because it has many people overlapping images, images with different angles and projections, different types of styles and poses images.

Approach

In order to select the appropriate approach for the problem, the paper first looks into few research approaches for object detection and localization, and then few approaches on cross-depiction detection. Both of them are reviewed individually.

For the object detection and localization, the sliding window approach didn’t turn out to be very efficient after the number of layers in the CNNs started increasing. Therefore few of the authors came up with better approaches. Few of the famous approaches discussed about object detection and localization included adding regression layer and even pooling layer in some cases to the CNNs and also including extra few CNN layers and outputting a separate vector for location of bounding box. Later on fast R-CNNs were introduced which outputted the location of bounding box, along with class detection score. This approach was picked for the ‘People-Art’ dataset. The fast R-CNNs were included with even more convolutional layers to make them even more fast and efficient. Lastly the YOLO approach was discussed where a single CNN is operated on an entire image and its divided to various rectangular cells where each cell outputs bounding box prediction and class probabilities simultaneously.

In the cross-depiction detection and matching research, a lot of work was done on matching the photograph images and hand drawn sketches. Colour channels were used to match different types of colour sketches, from more detailed coloured ones to roughly coloured sketches. Another approach was proposed where 2D sketches and 3D projections were compared and the CNN was optimized to reduce distance between 2D sketches and 3D projection for same class. An ample of work was done to perform cross-depiction detection in hand made sketches by finding various patterns. PAM( Part based model) was introduced for cross-depiction matching between hand made sketches and photographs which considered a fully connected graph and SSVM (Structured Support Vector Machine). They had separate attributes for both photographs and artwork where both the responses were observed and the indifference was calculated. This proved to be effective for detecting objects in hand-works and sketches.

The ‘People-Art’ requires more effort because of its complexity. Hence fine tuning of CNNs was considered which yielded better results.

Model

The model consists of two stages of CNN. In the first stage of CNN, the input is an image and rectangular region proposals created by many algorithms. It consists of many convolutional layers, max pooling layers, ReLU layers, and normalization layers in some cases. The final layer of this stage is ROI pooling layer which gets one input from previous convolutional layers and another input from regional proposal. The output is a feature vector with CHW dimensions where C is number of channels of previous convolutional layer.

In the second stage of CNN, the input is the output feature vector of previous stage CNN. This stage of CNN consists of max pool layers, ReLU layers and also some drop-out layers to avoid over-fitting of the model. The output is a set of co ordinates of the bounding box as well as score for each class. The final layer was modified to have output for only one class, that is person.

The fast R-CNN are trained by setting weights of first few layers from some pre-trained models. The fine tuning of the model for ‘People-Art’ dataset is done by using 3 models which each have same dimensions and 2 inner product layers followed by ReLU layers and drop-out layers.

Performance

The results whose Intersection over Union overlap(IoU) with the ground bounding box were above 0.5 was considered positive ROI and IoU in range [0.1,0.5) was considered negative. The ROI with IoU in range [0.4,0.6) was ignored because they were not giving clear results.

To increase the performance, the lower bound of negative ROI was removed so even the images with human similar classes like animals were included. This also helped the CNN to better distinguish between features with humans and without humans in artwork.

All the hyper-parameters for the CNN were fixed except the number of layers. The weights were fixed from other pre-trained models which were selected based on validation performance. It was observed that the performance was bad for number of layers more than 5. The first few layers were giving a good performance and were doing a good work in transferring images to artwork. The final features for the CNNs were selected based on the performance on validation sets.

It is seen that at higher threshold, the false positives were mainly caused due to poor localization and at lower threshold it was due to background regions. Comparatively, the higher number of false positives were due to poor localization. In some of the cases not complete picture of human was bounded and in few cases the CNN was confused due to the presence of more than one human in a single art.

Conclusion

It was observed that the fine tuning of CNN for photographs resulted in over-fitting which decreased the performance , while fine tuning of CNN on artworks yielded better performance. The best performance on ‘People-Art’ was found to be less than 60%. The dataset could also be improved by including artworks of even more different styles like Egyptian, Chinese, African etc which would widen the scope of this research.