Visual Probing and Correction of object recognition Models with Interactive user Feedback

Category: Research

ABSTRACT

With the advent of state-of-the-art machine learning and deep learning technologies, several industries are moving towards the field. Applications of such technologies are highly diverse ranging from natural language processing and computer vision to robotics. Object recognition is one such area in the computer vision domain. Although proven to perform with high accuracy, there are still areas where such models can be improved. This is in-fact highly important in real-world use cases like autonomous driving or cancer detection, that are highly sensitive and expect such technologies to have almost no uncertainties. In this paper, we attempt to visualise the uncertainties in object recognition models and propose a correction process via user feedback. We believe, this is highly effective as incorporating expert knowledge into a system that is mostly portrayed as a black-box can aid the performance of such models.

Index Terms: Information Visualisation, Uncertainty Visualization, Human Perception, Cognition

1 INTRODUCTION

Computer Vision technologies has made its way to almost every possible field. Image classification, classification and localisation, semantic segmentation are some of the tasks carried out in this area. Object recognition is one such task where given an image to the model, it can identify multiple classes/objects in the image but also localise each object by predicting the coordinates of the bounding box for each object. Although we can train the model with a huge data set, being able to generalise the model to cater to real world scenarios is highly challenging. Some of the issues that are faced by state-of-the-art object recognition models are:

- Real time detection: This challenge is with respect to processing/detection in high speed frames of images. It is mostly with respect to object recognition in a video stream. Handling this problem can be of relatively lower priority as there could scenarios where the users might expect some loss of frames during which the object recognition model might not be able to deliver the result.
- Handling multiple aspect ratios: This challenge is about addressing multiple scales of a given object in several inputs. This is one of the issue that we hypothesized our analysis and have proposed a solution.
- Limited data: This is one of the most challenging tasks. In order to train a model, we need lots of accurately annotated ground truth. Usually researchers leverage on crowd sourcing to gather data. One of the reasons that this might pose as problem is the reliability in the annotated results. The expert is usually not involved in the task of preparing the ground truth for training the object recognition models. This is another issue we have addressed in the paper. As the limited data is almost always the case, it makes sense to leverage expert user's knowledge about the most ambiguous cases in the limited data to increase the generality of the model.
- Skewed data: This problem is present in almost all problems. And it is quite difficult to get rid of as it is impossible to create a data-set that can cover all possible variances of a given problem at hand. The only way to address this may be is to regulate or decrease its impact.

As discussed above, it is cumbersome to create a data set that covers all possible variances that can occur in the data. This in-turn poses a problem in generalising the model to cater to new incoming test points. One way this can be addressed is to leverage expert knowledge to guide the model to handle special cases that may occur in the data. In this paper, we attempt to address this problem by:

- Proposing a generalised set of techniques that can be applied to any object recognition model.
- The techniques cover the aspects of analysing the current performance of the object recognition models, in-turn deciding factors that are leading to sub-optimal model performance.
- Incorporate user feedback in the process of correcting the classifier. This in-turn helps the classifier to regulate itself to handle similar ambiguous cases.
- Evaluate the proposed approach against the YOLO object recognition model, as part of the VAST 2020 IEEE Challenge.

2 RELATED WORK

As introduced earlier, our intention in analysing the performance of the object recognition model and proposing correction measures seeds from the involvement of expert user's perception and assessment. [2] is an approach that proposes finer improvement in the image classification model by engaging with the user to differentiate between the false positives and the true positives. Although this is one of the correction measures that we have proposed, in our case, we leverage the expert user's free will and rational thinking to decide on which objects in the images might have greater number of False Positives. In [2], based on certain criteria, they filter out the images which the classifier predicts with greater confidence and ask the user's opinions on those images. However, we leverage on selective attention where we probe the user to decide by himself on how likely an object will map to a False Positive. Also, we consider the case of multi-objects recognition in the image. [3] proposes a technique to allow the user to annotate objects in the images while incrementally training the classifier. However, their main focus area is improving the speed of the annotation process. In our paper, image re-annotation is one of the approaches we are proposing for classifier correction, but only for the images which have high degree of clutteredness. This paves way for improved generalisation of the object recognition model. In [1], they propose to improve the noisy labels in the training data by using a pretrained classifier to identify potential noisy labels in the images which are then presented to the user to allow him to re-tag the noisy labels. These corrections are then fed back to the training model. Although, our approach is on the similar lines of [1], where in based on a given state of the object recognition model, we allow the user to go about the correction process, the motive of [1] does not coincide with our purpose of incorporating user feedback and selective attention.

3 PROPOSED APPROACH

In comparison with a typical object recognition model, training images are utilised by the model for learning a pattern for each

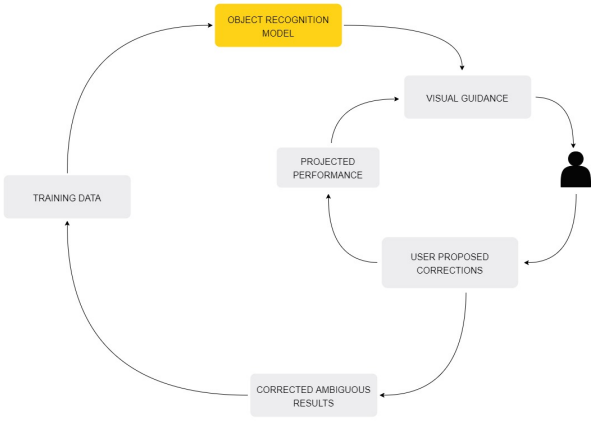

Figure 1: Proposed Approach

object. However, beyond this point there will be scenarios when the model would encounter a test point which is totally unfamiliar to the model, during which it is likely to fail if the input point is new and dissimilar to what was learnt by the model. We propose an approach in the lines of Online Learning where we would like the model to dynamically adapt to novel and ambiguous patterns. Hence, we propose an approach, where the Expert user is actively involved in the correction process of the object recognition model.

As seen in Figure 1, we include the Expert user in the introspection of the object recognition model. Typically the training data is given to the model which in-turn trains itself with the training images. These predictions are taken through a series of analysis steps to enable the user gain deeper insights into the model's performance and conclude how one can fix things. The user is also allowed to make corrections in the errors committed by the classifier. These changes in-turn affects the performance of the model. The projected performance of the classifier once the changes are made, are made available to the Expert user who can take decisions appropriately. The corrected data is then passed back to the model enabling it be better prepared for such ambiguities. We formulate two approaches for associating the model with the user.

- Classifier Analysis
- Classifier Correction

3.0.1 Classifier Analysis

In this phase, we try to evaluate the performance of classifier by introspecting its predictions. The predictions include the bounding boxes for each object in the input image and their confidence scores. We consider the following approaches for analysing the classifier performance:

- **Correlating Mean and Variances of the confidence scores for each object with its average bounding box size**. While it is expected that the objects with larger bounding sizes will have higher confidences, we can leverage this analysis to identify and isolate outlier points and conclude how or why such cases exist by analysing the corresponding images.
- **Correlating the degree of clutteredness of objects in the image with the average confidence score for that image**. We interpret this degree as the average density of objects in the image, and is computed as given in Eqn 1.

$$\rho = \frac{\text{\# objects predicted}}{(\max(xcoords) - \min(xcoords) \times (\max(ycoords) - \min(ycoords))} \quad (1)$$

where, ρ is the average density of objects in the a given image, *xcoords* represents the x-coordinates of the bounding boxes of objects and *ycoords* represents the y-coordinates of the bounding boxes of objects in that image.

- An overview of the **average confidence scores for each object** that in-turn enables us to isolate the objects that were likely to have been misclassified is also presented to the user.

We present the above analysis as visual information that can be easily perceived by the user. This in-turn will aid the user in asking the right questions to rectify the model.

3.0.2 Classifier Correction

In recent times, Computer Vision models are tuned and tweaked to match their perceptive skills with that of a human. Rather than just manipulating the models' internal workings, in our approach, we considered the perception of the user himself while improving the models. This should bring more clarity for the models during their predictions on the ambiguous scenarios. We consider the following correction mechanisms:

- **Eliminating False Positives through user feedback**. It is common that some objects will be better predicted by the classifier than others. This is purely based on how accurately the classifier differentiates the objects. Hence, it is quite often that there will be existence of a lot of False Positives in the data - which are objects in the images that were wrongly classified. eg: An object that is circular, orange in color has a high probability of it being an Orange. At the same time, if the model was also given a ball that is orange in color, this leads to creation of False positives.

  Object recognition models use a series convolutional layers to learn features/intricate patterns of various objects. If such patterns are similar in two objects, this could be misleading to the classifier. One way to address this issue is to include the Expert user to assess and isolate such False Positives. The user, through his perception concludes that object in a give image was confused with a different object. There could be several reasons as to why the user arrived at this conclusion. One of the reason could be, an object's patterns highly matches the features of another object's training images. The model could then be informed about the existence of such False positives that leads to an improved performance.

- As discussed earlier, an image with high degree of clutteredness i.e, high average density of objects will have a poorer confidence score by the model. This results in inaccurate predictions of the bounding boxes. In such cases, unless there exists training examples for the same image with different scales of objects, it is quite difficult for the classifier to learn the patterns for objects that are highly cluttered. This is especially a problem if the training data is very limited.

  In such cases, we allow the Expert user to intervene and **re-annotate the ambiguous images** which in-turn improves the performance of the classifier with respect to such images where the objects are highly cluttered.

  In all the above techniques, the common and one of the important aspects is the inclusion of user Feedback while improving the model's performance to make it more human like.

4 Evaluation on YOLO object recognition model

We apply and evaluate our proposed approaches to analyse the performance of YOLO object recognition model. The predictions of the model (bounding box co-ordinates and the classes) were provided by the VAST 2020 Mini Challenge.

4.0.1 Our Tool

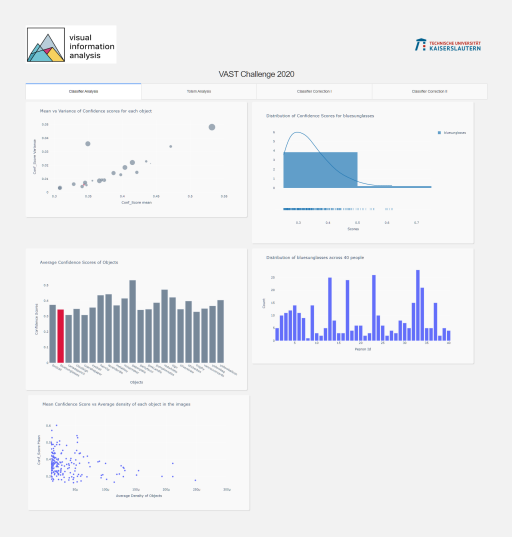

Figure 2: Overview of our Tool

As seen in Figure 2, we attempted to provide a simple yet interactive UI to the user. The graphs that can be seen in the tool are interactive and dependent on each other. We used Plotly's DASH framework to develop the web-application. DASH provides us with out-of-the-box APIs for setting up callback events to make the graphs interactive and interdependent.

4.0.2 Analysis

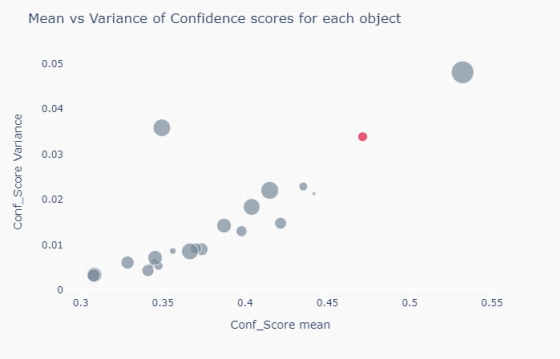

Figure 3: Mean and Variance of Confidence Scores vs the Average Bounding Box size for each object

- As seen in Figure 3, it depicts how the bounding boxes affects the average confidence scores for each object that was recognised by the classifier. It is expected and can be concluded from the graph that the objects with higher average bounding box sizes have higher average confidence scores. However, the points that do not adhere to this expectation might be of interest and might require further analysis. Hence, one of the suggested steps to improve the performance of the object recognition model could be to have training images with larger sized objects so that the model learns the intricacies of the object's texture.

- As per Figure 4 a brief high level overview can be obtained regarding which objects were predicted with highest confidence

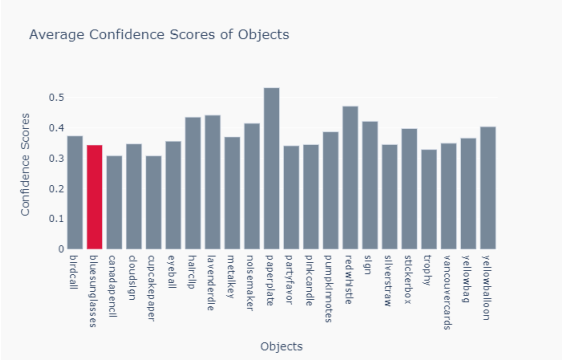

Figure 4: Average Confidence Scores for each predicted object

scores. This insight in-turn will help in the correction process discussed in the next section.

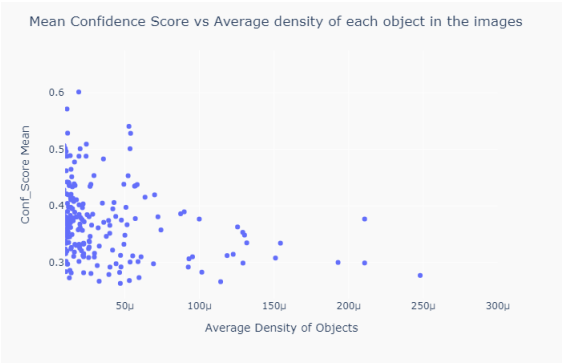

Figure 5: Average Confidence scores of images verses their average bounding box densities

- As per the Figure 5 depicts how the average confidence scores of images vary with the average densities of objects in each image. The density ρ is computed as seen in Eqn 1. It is seen that there is a general negative correlation between the confidence scores and the average densities of the objects. This might imply that the confidence scores of objects for images with objects that are very close to each other are relatively lesser compared to the rest. The YOLO models do not generally perform well for objects that are cluttered which in-turn is realised by our insight in the Figure 5.

4.0.3 Correction

- As the first correction process, we enabled the user inspect the images for which a given object was predicted. This was then assessed by the user to determine its validity and hence treat them as a True or a False Positive. To smoothen the whole process further, we introduced selective attention by guiding the user through the process. This was done by showing the user, the objects that are more likely to contain False Positive images using a graph that portrayed proportions of the test images mapped to each object. As seen in the Figure 6 , the object "eyeball" is likely to contain more False Positives - as it was mapped to the test images maximum number of times. This kind of analysis is even more effectively used, if the user is aware of the how many objects are mapped to each of the test images.

  Once, the user selects the images to inspect, it is displayed as a grid which in-turn can be selected by the user for elimination.

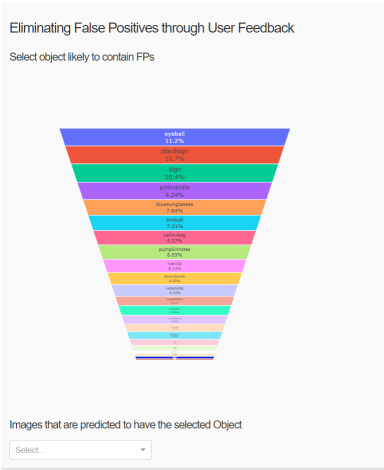

Figure 6: Selective Attention

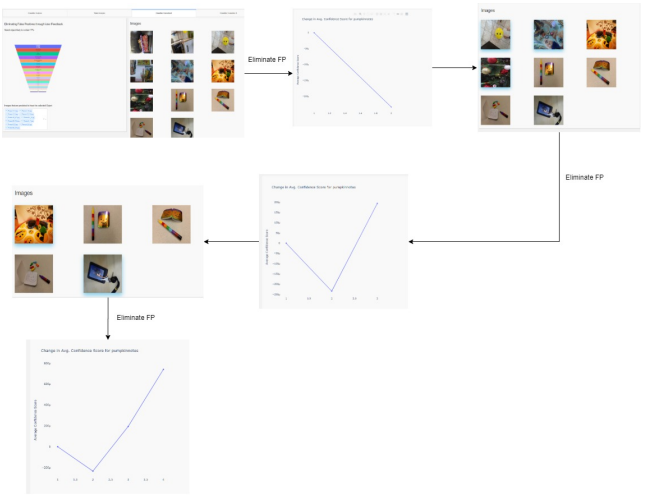

Figure 7: False Positives Elimination Process

These selected False Positives are treated as flawed results by the classifier, and hence we project how the performance of the classifier improved in terms of the average confidence score for that object. This projection graph is aggregated with its previous changes, and hence gives an overall sense of improvement in the model. This also decreases uncertainty in the model with respect to the object. The entire flow can be seen in Figure 7. As seen, initially the user chooses a set of images that was mapped to an object. This is followed by successive selections and eliminations of False Positives. During each stage of elimination, a graph records how the performance of the object recognition model with respect to the object is being improved.

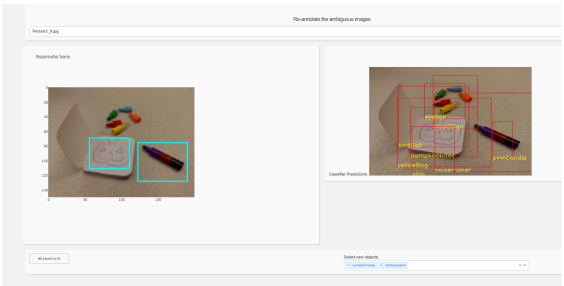

Figure 8: Re-annotating ambiguous images

- For true positives images, which are the ones that are the true expected object predictions, there are possibilities that the bounding boxes of the objects were not accurate enough. This can seen in the analysis Figure 5 where highly cluttered objects aren't accurately predicted. Hence, we propose an approach as seen in Figure 8 where we allow the user to re-annotate the objects by showing the actual predictions of the bounding boxes to the user. The re-annotated co-ordinates can then be exported by the user that can in-turn be used to update the training model to improve its performance.

5 Future Work

An image is composed of distribution of pixel intensities in the red, green and blue channels. An image with high contrast between foreground and background will roughly have strong non-uniform distribution of the pixels in the RGB channels. On the other hand, less contrastive images will have uniform distributions. A suitable thresholding technique like OTSU Thresholding [4] can be explored here. Hence we intend to conduct further analysis on these lines by correlating the pixel distribution of the RGB channels of an image with how well the objects in the image were predicted by the object recognition model.

We also intend to learn in an unsupervised way the clusters that could be obtained from the test images. Test images could be subject to dimensionality reduction techniques like PCA, t-SNE, etc and this reduced representation could be clustered using techniques like the KMeans algorithm. These clusters can then be correlated with the confidence scores of the objects in the test images to draw further conclusions about the object recognition model.

6 Conclusion

In this paper, we explored ways of improving the performance of object recognition models. We proposed a set of approaches in the lines of classifier analysis and corrections. We also explored ways of leveraging user perception to humanize the model's perception. These approaches were implemented as part of the VAST 2020 Challenge and respective results were presented.

References

[1] A. Bäuerle, H. Neumann, and T. Ropinski. Classifier-guided visual correction of noisy labels for image classification tasks, 2018.

[2] Y. Cui, F. Zhou, Y. Lin, and S. Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1153–1162, 2016.

[3] A. Yao, J. Gall, C. Leistner, and L. Van Gool. Interactive object detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3242–3249, 2012.

[4] C. Yu, C. Dian-ren, L. Yang, and C. Lei. Otsu's thresholding method based on gray level-gradient two-dimensional histogram. In *Proceedings of the 2nd International Asia Conference on Informatics in Control, Automation and Robotics - Volume 3*, CAR'10, p. 282–285. IEEE Press, 2010.