"Unveiling the Invisible" – Deep Learning-based Semantic Segmentation for Analysing Activity Patterns

Gurkiran Kaur and Li Zhang

Department of Computer Science, Royal Holloway, University of London
Surrey, TW20 0EX, UK
E-mail: zkac303@live.rhul.ac.uk; li.zhang@rhul.ac.uk

The ubiquity of internet-enabled devices has led to a rapid increase in the use of connected cameras for real-time monitoring, creating a high demand for (automated) visual data analytics across various industries. The prospect of automating visual data analysis to drive positive change involves extracting actionable insights from data that will inform decision-making processes, improving efficiency, and contributing to evidence-based strategies across diverse applications and industries. This research explores and compares well-known semantic segmentation models such as DeepLabV3+ and UNet, determining the best-suited for use in a visual analytics and scene understanding, culminating in a proof of concept program capable of automating video analysis, plotting detections, average trajectories, and identifying outliers.

Keywords: DeepLabV3+; Automated Data Analysis; Semantic Segmentation.

1. Introduction

Analysis of activity patterns within a scene using object tracking and visual data analytics have long been the focus of extensive research and development efforts, as evidenced by the multitude of seminal works across a range of industries [1]. These techniques have important applications across a wide range of industries and scenarios, from surveillance and autonomous vehicles to healthcare in the automated identification of cancerous tumors. The ability to monitor, analyze, and interpret visual data streams has revolutionized decision-making processes and driven advancements across diverse domains.

The research delves into the complex technical landscape of these technologies, exploring cutting-edge methods, models, and tools that fuel the nexus between computer vision and data analytics offering insight into fundamental components that will guide development of our own proof of concept tool capable of creating modal average activity masks for a given scene.

The project aims to automate and streamline visual data analysis, addressing the tedious and error-prone nature of the task, while offering potential solutions

for various industries such as policing, surveillance, research, and traffic management, where large volumes of video data require efficient and accurate processing.

2. Related Work

The research holds significance due to its unique aims, particularly in the context of a scarcity of directly comparable endeavors, though this also presents challenges in terms of available support. Franklin et al. [2] demonstrated a highly effective anomaly detection method in video footage, closely aligned with our proposed approach, utilizing a Convolutional Neural Network (CNN) and the COCO dataset for real-time video data. Similarly, Bajestani et al. [3] presented a system for anomalous vehicle and pedestrian detection using object recognition techniques, contrasting with our proposed higher granularity semantic segmentation techniques. Santhosh et al. [5] conducted a comprehensive survey of existing literature, offering valuable insights into contextual anomalies and statistical methods for quantifying deviation, which will guide our research and serve as a reference point for industry standards and common practices.

3. Preliminaries

Remark 1. An anomaly is a data point which does not fit the patterns of normal behavior, which will be defined by our average heat map created from multiple segmented images of a scene over time.

Remark 2. Normal activity encompasses data points that conform to the established patterns of behavior as defined by the average heat map derived from multiple segmented images of a scene over time. This definition ensures that normal activity is reflective of the prevailing spatio-temporal patterns detected by the system.

Remark 3. Conformity is exhibited by data points or objects that align with the established patterns of behavior, as defined by the average heat map derived from multiple segmented images of a scene over time.

4. Methodology

4.1. *Proposed Architecture*

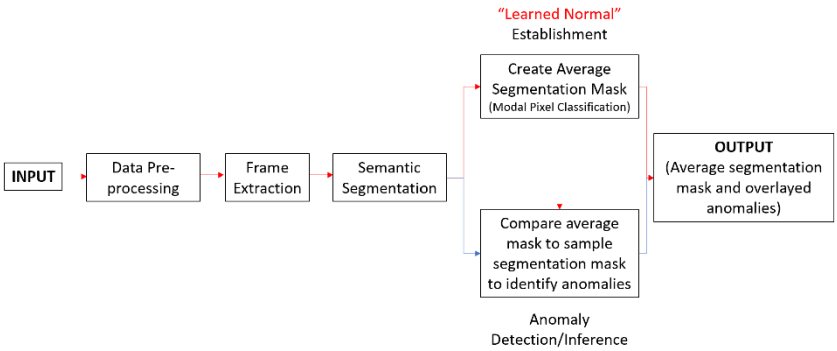

Fig. 1. The proposed pipeline structure of the visual data analytics tool

Table 1. Table describing the broad function of each described section in the pipeline architecture proposed in Fig 1.

Stage	Function
Input	Acquire initial raw data. The data should be from a single fixed camera and should be at least 30 seconds in length.
Data Pre-processing	Raw data should be standardized, addressing inconsistencies or errors in data.
Frame Extraction	Reduce the number of frames to reduce computational burden. Extract frames at timed intervals.
Semantic Segmentation	Create multiple segmentation masks for each given frame.
Average Heatmap ("Learned Normal")	Create average heatmap from the modal classification of each pixel within the segmented frames. This will be used as a reference point for normal activity.
Anomaly Detection	Compare newly segmented frames to our 'learned normal' and identify potential outliers. Create a 'delta map' showing the differences in expected normal to the newly segmented scene.
Output	Output desired results via our average heatmap along with potential anomalies, that could be marked onto frames.

4.2. *Semantic Segmentation Model Comparison and Results: DeepLabV3+ vs UNet*

Both DeepLabV3+ and UNet were chosen for their prominence in semantic segmentation tasks and as models that continue to show high efficiency and

promise in innovative machine learning applications. DeepLabV3+ employs a deep CNN with atrous convolutions to capture multi-scale contextual information, while UNet features a U-shaped encoder-decoder architecture for high-resolution feature extraction, commonly used in biomedical research.

4.2.1. *UNet*

Our U-Net Jupyter Notebook file documents our experimentation in creating a U-Net model, following resources and code from the pyimagesearch online tutorial [6], and trained on the 'Oxford iiit pet/3.2.0' dataset [7]. Despite the seeming irrelevance of this dataset to our use case, this experimentation serves solely for model comparison purposes, allowing us to assess the average performance of our chosen models, DeepLabV3+ and UNet, using a dataset of comparable size to those we anticipate using in our program.
After training for a suggested 20 epochs, the final statistics on the UNet model stood as follows:

- loss: 0.2892
- accuracy: 0.8884
- val loss: 0.339029
- val accuracy: 0.8757

A loss value of 0.2892 suggests good performance, while the 0.89 accuracy value indicates strong performance. Working on unseen data, or the validation set, a loss of ~0.34 suggests good performance on unseen data and as for accuracy, at ~0.88, this implies that the model predicts outcomes correctly about 88% of the time, which is an impressive result.

Fig. 2. Loss and Accuracy Graphs for both Validation and Training Data for UNet model

The graphs in Fig. 2 show expected curves whereby our accuracy increases as our loss decreases. Furthermore, the validation curve closely follows the training curve, indicating that the model performs well on unseen data (generalizes well) without overfitting.

4.2.2. *DeepLabV3+*

Our DeepLabV3+ Jupyter Notebook documents the creation and training of our DeepLabV3+ model, following steps and resources from the Keras website tutorial [8]. Differing from the UNet approach, we trained our DeepLabV3+ model on the Crowd Instance-level Human Parsing Dataset (CIHP) [9] due to its specific data input requirements.

After training for 20 epochs, as we did with our UNet model, these are the statistics we obtained:

- loss: 0.2214
- accuracy: 0.9339
- val loss: 1.0525
- val accuracy: 0.7517

A loss value of 0.2214 suggests good performance, while the 0.93 accuracy value indicates very good performance, suggesting that the model is correct around 93% of the time. The model's performance on the validation set, with a loss of 1.0525 and an accuracy of 0.75, suggests there is potential for improvement, possibly due to overfitting. Further optimization and regularization techniques could enhance the model's generalization ability and address these issues.

Fig. 3. Loss and Accuracy Graphs for both Validation and Training Data for DeepLabV3+ model

The graphs in Fig. 3 show expected curves whereby our accuracy increases as our loss decreases, however, there is a slight discrepancy between the validation and training set curves. This is likely attributable to model overfitting or a mismatch in dataset and model complexity. This should not be a cause for alarm as we will be using a different implementation and dataset in our tool. A good example of the output of our DeepLabV3+ model performing on unseen data

can be seen in Fig. 4, which depicts the original image, alongside the overlaid mask and the mask alone.

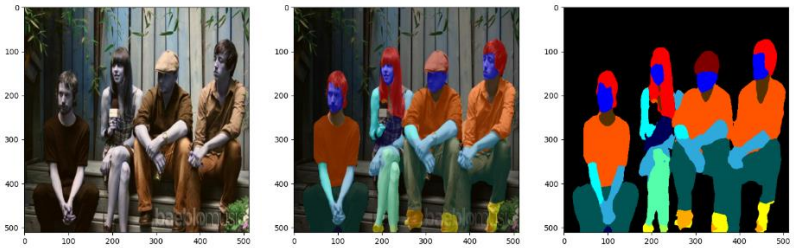

Fig. 4: An example output of running the DeepLabV3+ model on an unseen image

4.2.3. *Summary*

DeepLabV3+ demonstrated superior performance with a final accuracy of 0.9339 and validation accuracy of 0.7517, outperforming UNet's final accuracy of 0.8884 and validation accuracy of 0.8757 after 20 epochs. DeepLabV3+ also exhibited lower validation loss, indicating better generalization. While acknowledging inherent complexities limiting perfect performance, such as label ambiguity and computational constraints, DeepLabV3+ remains the preferred model for advancing our project despite slight divergences in training and validation curves.

4.3. *Implementation*

We utilised the Pixellib [10] library for efficient and reliable implementation of our DeepLabV3+ model for semantic segmentation. Pixellib simplifies the integration of complex machine learning technologies into software solutions, addressing our project's challenge of limited time and technical expertise. Unlike other libraries, Pixellib leverages deep learning techniques like DeepLabV3+, making it suitable for our pipeline. It also supports future work development by facilitating the training of custom segmentation models and enabling real-time video segmentation. Our implementation uses the DeepLabV3+ model trained on the ADE20K dataset, which includes classes relevant to our project such as vehicles and pedestrians.

Fig. 5: A TfL JamCam frame with and without the Pixellib segmentation mask

Utilizing the Pixellib library for semantic segmentation, we concentrate on generating average activity maps to pinpoint anomalous and conformal activity within scenes. By creating segmentation masks from frames at regular intervals, we derive an average activity heatmap (using the modal pixel classification), aiding in the identification of normal and anomalous behavior. Figs 5 and 6 shows more example segmentation outputs using DeepLabV3+.

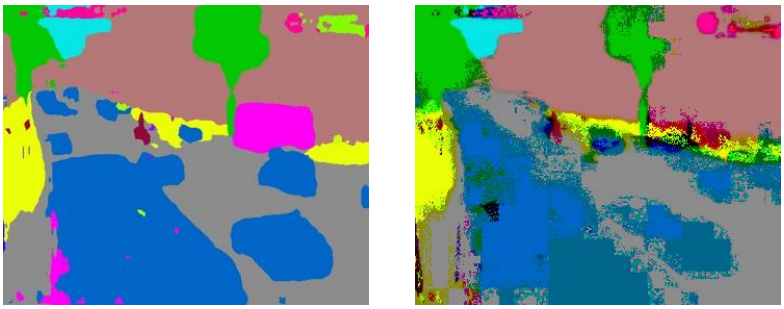

Fig. 6: (Left) Segmentation mask created using Pixellib DeepLabV3+ of a frame in the video referenced in above (Fig. 5), and (Right) Average activity mask generated from a collection of frame segmentation masks (like the one on the left) from our pipeline using DeepLabv3+ semantic segmentation. Judd Street (West) Transport for London JamCam

5. Conclusion

This research marks a successful exploration into novel computer vision techniques for automating decision-making and visual analysis, representing a valuable learning opportunity in my pursuit of a career in machine learning and mathematics. The creation of average activity masks demonstrates technical success, meeting our initial aims for a proof of

concept tool using semantic segmentation. Moving forward, there is potential for further exploration, including delving into deep learning methodologies [11-13], researching data sanitization's impact, and exploring novel trajectory plotting methods.

References

1. E.R. Davies, Computer and machine vision: theory, algorithms, practicalities. Amsterdam; Boston: Elsevier, 4th ed., 2012.
2. R. J. Franklin, Mohana, and V. Dabbagol. Anomaly detection in videos for video surveillance applications using neural networks, in International Conference on Inventive Systems and Control, pp. 632–637, 2020.
3. M. F. Bajestani, S. S. H. R. Abadi, S. M. D. Fard, and R. Khodadadeh, AAD: Adaptive Anomaly Detection through traffic surveillance videos, Aug. 2018. arXiv:1808.10044 [cs].
4. S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Jan. 2016. arXiv:1506.01497 [cs].
5. K. K. Santhosh, D. P. Dogra, and P. P. Roy, Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey, ACM Computing Surveys, vol. 53, pp. 119:1–119:26, Dec. 2020.
6. M. Maynard-Reid, U-Net Image Segmentation in Keras, Feb. 2022.
7. Visual Geometry Group - University of Oxford. Dataset containing around 8000 images of dogs and cats.
8. K. Team, Keras documentation: Multiclass semantic segmentation using DeepLabV3+. Tutorial used for using DeepLabV3+ to Segment Images.
9. Papers with Code - CIHP Dataset, *paperswithcode.com*. https://paperswithcode.com/dataset/cihp
10. Papers with Code - Simplifying Object Segmentation with PixelLib Library. https://paperswithcode.com/paper/simplifying-object-segmentation-with-pixellib
11. S. Slade, L. Zhang, L., H. Huang, et al., (In Press). Neural Inference Search for Multiloss Segmentation Models. *IEEE Transactions on Neural Networks and Learning Systems*.
12. L. Zhang, S. Slade, C.P. Lim, et al., (2023). Semantic segmentation using Firefly Algorithm-based evolving ensemble deep neural networks. *Knowledge-Based Systems*, 277, p.110828.
13. L. Zhang and C.P. Lim, 2020. Intelligent optic disc segmentation using improved particle swarm optimization and evolving ensemble models. *Applied Soft Computing*, 92, p.106328.