iJournals: International Journal of Software & Hardware Research in Engineering
ISSN:2347-4890, ijournals.in/ijshre
Volume 13 Issue 1, January 2025

A Smart Crime Reporting Bot Using YOLO-Based Weapon Detection And RNN-Based Text Analysis

Author: Seowoo Choi

Seoul International School, Seoul, Republic of Korea

seowoochoi30@gmail.com

Abstract – Home invasion crimes in South Korea, though relatively rare, still occur, particularly in urban areas like Seoul, where property crimes such as burglary and robbery are more prevalent. Despite low violent crime rates, the rise in property crimes has prompted many households to adopt advanced security systems, including CCTV cameras. The police respond quickly to such incidents, often using surveillance footage and data analysis, although the lack of physical evidence or witnesses can complicate investigations. This paper proposes a smart security system that combines YOLO for real-time weapon detection and RNN-family models (RNN, LSTM, GRU) for processing emergency messages. Evaluation results demonstrate the system's effectiveness in detecting weapons, tracking intruders, and generating timely reports, showcasing the potential of deep learning techniques to enhance home security. By integrating advanced object detection and message analysis, the proposed system offers a promising solution to improve response times and reduce the risks associated with home invasions.

Keywords: Crimes, Security systems, Image detection, Deep learning, RNN, YOLO, NLP

1. Introduction

In South Korea, home invasion crimes are relatively rare, but they still occur, particularly in urban areas. These crimes often involve individuals unlawfully entering homes, either for theft or with the intent to cause harm. While violent home invasions are less common compared to other countries, South Korea has seen a rise in property crimes, including burglary and robbery, especially in densely populated areas like Seoul. Home invasion cases that involve violence, such as intruders threatening or harming residents, have been reported, though these remain infrequent.

Fig. 1

Despite the relatively low rates of violent crime, home invasions remain a concern for South Koreans, particularly with the increase in urbanization and the challenges posed by crowded living conditions. In response to these threats, many households have turned to advanced security systems. The installation of CCTV cameras, doorbell cameras, and smart security devices has become increasingly popular, as people seek ways to protect their homes and families.

The police in South Korea generally respond quickly to crimes, including home invasions, which are taken very seriously. Investigations are often aided by surveillance footage and high-tech data analysis methods. However, as with any crime, the lack of physical evidence or witnesses can complicate the resolution of such cases. The authorities in South Korea maintain strict laws to deter criminal activity, with severe penalties for home invasion and related crimes, especially when violence or weapons are involved.

Overall, while home invasions are not as frequent in South Korea as in some other countries, the increasing use of technology for security and the rigorous enforcement of laws play a crucial role in reducing their prevalence and ensuring public safety.

Current systems are also prone to false alarms and leave uncovered areas vulnerable. Motion sensors, which typically cover only small areas, can miss critical spots even when multiple sensors are used. Additionally, false

alarms caused by weather, animals, or passers-by are common. One significant drawback of motion sensors is the lack of evidence to identify the intruder. This limitation can be addressed by integrating video recording and analysis. The declining prices of cameras make them more accessible to low-income households.

The main contribution points of this paper are: 1) detect the weapon based on YOLO; 2) analyze the reporting message from the victims based on RNN family; 3) proposing a smart system using former methods.

In chapter 2, the image classification methods and the NLP methods are introduced. In chapter 3, the proposed system is demonstrated. Evaluations are in chapter 4. The paper concludes with future works in chapters 5 and 6.

2. Background

2.1 Image Classification Methods

The human visual system is remarkably fast and accurate, enabling individuals to intuitively recognize objects, understand their spatial relationships, and perform complex tasks, such as driving, without conscious thought. This efficiency is mirrored in the development of object detection algorithms, which enable computers to perform tasks like autonomous driving, real-time assistive technologies, and robotic systems. Fast and accurate algorithms for object detection allow for such applications without relying on specialized sensors.

2.1.1 CNN

Convolutional neural networks (CNNs) are a class of artificial neural networks that have become highly popular for image analysis. CNNs are particularly adept at detecting patterns within images, which makes them invaluable for tasks such as object recognition. The network's layers, known as convolutional layers, process input data by applying convolutional operations to extract and transform features, which are then passed to the next layer. These convolutional layers can detect simple patterns, such as edges and textures, in early stages and more complex features, such as faces or people, in deeper layers. The performance of CNNs improves as the network deepens, with filters becoming increasingly sophisticated in their ability to detect various objects.

2.1.2 RCNN

Prior to CNNs, image detection systems like Deformable Parts Models (DPM) and Region-based CNNs (RCNN) were used for object detection. DPMs employ a sliding

window approach, where classifiers operate at fixed intervals across an image to identify objects. RCNN enhances this by generating potential bounding boxes first, and then running classifiers on these boxes. The system then refines the bounding boxes through post-processing, eliminating redundant or similar boxes based on characteristics such as texture or color. However, RCNN requires multiple training phases and complex pipelines, making it slow and difficult to optimize.

2.1.3 SPP Net

Spatial Pyramid Pooling (SPP Net) is an improvement over RCNN, designed to address issues such as image warping. The process involves inputting a full image into a trained CNN to extract a feature map. SPP Net employs Selective Search to create region proposals of varying sizes, followed by feature extraction and processing through fully connected layers. The network then uses these features to train a binary classifier and bounding box regressor. The main advantage of SPP Net is that it requires only a single CNN calculation, significantly speeding up training and testing compared to RCNN.

The process begins by inputting the complete image into a trained CNN to extract a feature map. Each Region of Interest (RoI) produced by Selective Search varies in size and ratio, so Spatial Pyramid Pooling (SPP) is applied to extract a feature vector with a constrained size. This feature vector is then passed through fully connected layers for further processing. The extracted feature vector is used to train a binary Support Vector (SV) classifier for each image class, enabling the identification of objects within the image. Additionally, the feature vector is utilized again to train a bounding box regressor, refining the localization of detected objects within the image.

SPP offers several advantages over traditional methods like RCNN. By utilizing SPP, we can eliminate the warping and distortion issues inherent in RCNN, ensuring more accurate feature extraction. Additionally, while RCNN requires up to 200 CNN calculations for processing, SPP Net streamlines the process by requiring only a single CNN calculation. This significant reduction in computational demand leads to faster training and testing times, making SPP Net more efficient and time-effective in object detection tasks.

2.1.4 OpenCV

OpenCV (Open Source Computer Vision Library) is a comprehensive library that provides tools for image and video processing, including object and face detection. Its ability to capture and store video makes it invaluable for security applications, such as monitoring for intruders and

storing video evidence for later analysis. OpenCV can help identify human body parts, vehicles, signage, and other objects in video feeds, facilitating real-time surveillance and response.

2.1.5 YOLO

YOLO (You Only Look Once) is a widely used deep learning method for real-time object detection. Unlike traditional object detection techniques, which process an image multiple times to identify objects, YOLO treats object detection as a single regression problem. This approach allows YOLO to predict both the bounding boxes and class probabilities of objects in a single pass, making it faster and more efficient compared to older methods like R- CNN that require several stages of processing.

The core idea behind YOLO is to divide the input image into a grid, typically of size S×SS \times SS×S. Each grid cell is responsible for predicting a set of bounding boxes and the associated confidence scores. If the center of an object falls within a grid cell, that cell is tasked with detecting and classifying the object. YOLO then predicts the center coordinates (x, y), width and height (w, h) of each bounding box, as well as the confidence score that the box contains an object and its accuracy. Additionally, each grid cell predicts class

probabilities for the object, indicating which category the object belongs to (e.g., car, dog, person).

YOLO's method of detecting objects is framed as a regression problem. The network directly predicts the bounding box coordinates and class probabilities for each grid cell in one step, rather than using multiple stages for classification and localization. This regression approach reduces the complexity of the pipeline, allowing YOLO to perform object detection faster than other methods that rely on separate steps for each task.

One of the major advantages of YOLO is its speed. Since the model processes the entire image in a single pass, it is capable of performing real-time object detection, which is ideal for applications such as video surveillance or self- driving cars. YOLO also has the benefit of considering global context in its predictions, meaning it evaluates the entire image at once rather than focusing on individual regions. This leads to fewer background errors and helps improve object detection accuracy.

However, YOLO is not without its limitations. While the speed is an advantage, the localization accuracy can suffer, particularly for small objects in large images. Since the image is divided into a fixed grid, smaller objects may be detected less accurately due to the limitations of the grid size. Furthermore, although YOLO is fast, its precision may

be lower compared to other more complex models, such as Faster R-CNN or RetinaNet, especially in terms of fine- grained localization.

Since its introduction, YOLO has evolved through several versions. YOLOv2, also known as Darknet-19, introduced improvements in both speed and accuracy, allowing the model to generalize better across various datasets. YOLOv3 further enhanced the model by improving the detection of smaller objects and refining the architecture. YOLOv4 continued the trend of optimizing both speed and accuracy, incorporating better backbone networks and data augmentation techniques. Although YOLOv5 is not an official version, it was developed by the community with the goal of achieving better performance and easier deployment.

In conclusion, YOLO remains one of the most efficient deep learning methods for real-time object detection. Its ability to process images quickly and consider the global context of objects makes it particularly suitable for applications in surveillance, autonomous vehicles, and robotics. Despite its limitations with smaller objects and lower precision compared to some other models, YOLO's speed and versatility make it a leading choice for many practical use cases.

2. Natural Language Process methods

2.2.1 RNN

Recurrent Neural Networks (RNNs) are a class of artificial neural networks specifically designed to process sequential data, making them highly effective for tasks such as time series analysis, natural language processing, and speech recognition (cite). Unlike traditional feedforward networks, RNNs feature connections that loop back on themselves, allowing them to retain memory of previous inputs within a sequence. This inherent capability enables RNNs to capture temporal dependencies and patterns over time, rendering them particularly well-suited for tasks where the context of prior data points is essential for interpreting the current input.

Despite their advantages, RNNs face certain challenges, notably the issues of vanishing and exploding gradients during training, which can impede their performance when handling long sequences. To mitigate these challenges, variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been developed. These modifications enhance the ability of RNNs to effectively learn and retain long-term dependencies, addressing the limitations associated with traditional RNN architectures.
In summary, RNNs are fundamental in various applications requiring the processing of sequential data,

and their effectiveness is further improved through advanced architectures like LSTM and GRU, enabling them to address complex problems in fields such as natural language processing and speech recognition.

2. LSTM

Long Short-Term Memory (LSTM) networks are a specialized variant of recurrent neural networks (RNNs) developed to address the limitations of traditional RNNs, particularly the vanishing gradient problem that often arises when training on long sequences (cite). LSTMs mitigate this issue through a distinctive architecture that incorporates memory cells and three critical gates: the input gate, the forget gate, and the output gate. These gates control the flow of information, enabling the network to determine which data to retain, which to discard, and when to produce relevant outputs.

This architectural innovation allows LSTMs to effectively capture long-range dependencies within data, making them particularly adept at learning from sequences where prior contextual information is essential. Consequently, LSTMs are extensively utilized in applications such as natural language processing, speech recognition, and time series forecasting, where understanding and preserving temporal relationships within data is crucial for accurate predictions and decision- making.

3. GRU

Gated Recurrent Units (GRUs) are a variant of recurrent neural networks (RNNs) developed to capture dependencies in sequential data while addressing some of the limitations associated with traditional RNNs, such as the vanishing gradient problem. GRUs simplify the architecture of Long Short-Term Memory (LSTM) networks by consolidating the input and forget gates into a single update gate and introducing a reset gate that governs the extent to which past information is retained. This streamlined structure reduces model complexity, yet it remains effective in preserving relevant information over long sequences.

Due to their ability to learn from data where contextual relationships and temporal order are significant, GRUs are well-suited for tasks such as natural language processing, speech recognition, and time series analysis. Their computational efficiency and strong performance make GRUs a preferred choice for various applications requiring the handling of sequential data.

3. Proposed system(methods)

The proposed system of this research is demonstrated below in Fig.2. The system starts the process once the sensor detects some movement at the entrance. It ends when nothing is sensed in a pre-set time interval.

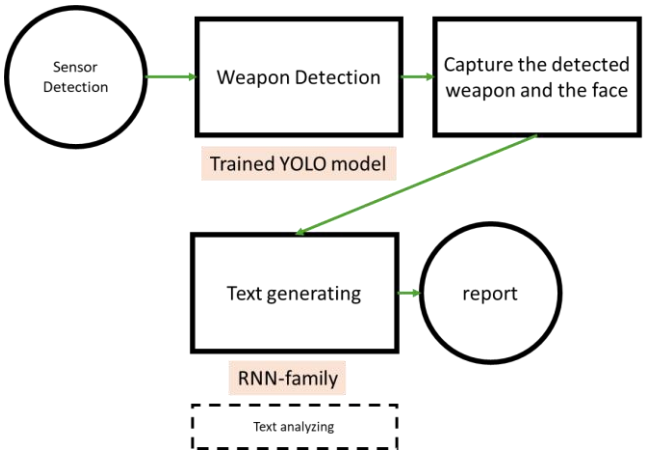

Fig. 2

1) with the trained YOLO model, the system detects a weapon such as knife, gun, or other tools which have potential to be weapons via the history.
2) One the system detects weapons; it records the weapon and the face and the body of the criminal. In this process even if the criminal covers their face, we can still estimate heights and silhouettes of him/her.
3) Using RNN-family, the system automatically generates text message with keywords of weapon and time with pre- stored location.
4) To filter whether the message is trick or real-emergency, the text analyzing is executed with trained RNN-family. In this step, the local database is used.
5) The final step is to report the text message which is generated and manually written by user to the nearest neighbor facilities.

4. Evaluations

All Experiments are conducted in Python with core i5-9gen. The models are trained with open dataset (YOLO) and the locally generated dataset (RNN family).

4.1 Image detection with yolo

The size of the images are refers that how far (clear) the image is. The table is showing percentage of success in detecting depending on the accuracy.

	1%	10%	30%	50%	70%
1KB	0	0	0	0.09	0.124
10KB	0	0	0	0.118	0.123
100KB	0.5	0.607	0.728	0.79	0.813
1000KB	0.630	0.699	0.875	0.905	0.987
10000KB	0.605	0.877	0.943	0.996	0.997

Table. 1

4.2 RNN-family Evaluations

RNN, LSTM and GRU has been evaluated with same train and validation dataset in this paper.

4.2.1 RNN

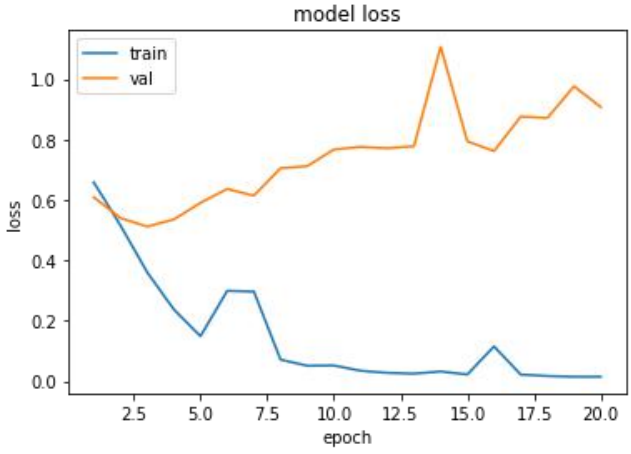

Fig 3

4.2.2 LSTM

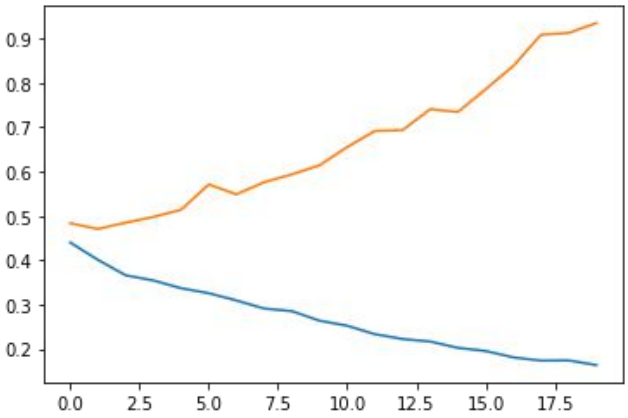

Fig 4

4.2.3 GRU

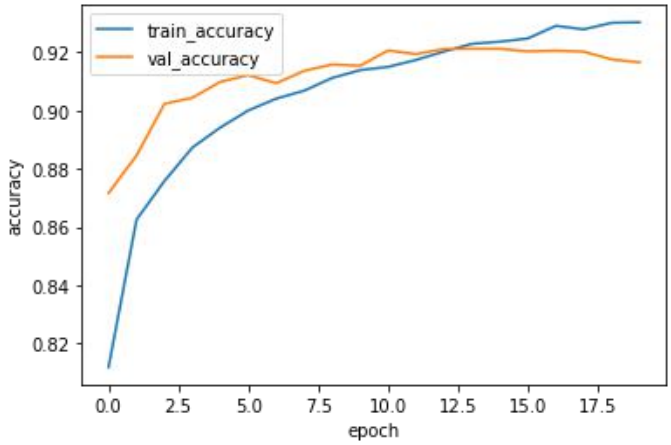

Fig 5

5. Conclusion

In conclusion, this paper addresses the growing concern of home invasion crimes in South Korea, specifically focusing on improving security systems using advanced deep learning techniques. The proposed system integrates YOLO for weapon detection, which allows for real-time identification of potential threats, such as knives or guns, based on historical data. The use of RNN-family models, including RNN, LSTM, and GRU, enhances the system's ability to process textual information generated from surveillance data, providing a more accurate understanding of the situation. This combination of object detection and natural language processing contributes to a comprehensive smart security system capable of mitigating the risks of home invasions and improving response times.

The evaluation results demonstrate the efficacy of the proposed system, with YOLO performing well in object detection tasks, even with varying image sizes, while the RNN-family models showcased their effectiveness in analyzing and filtering emergency messages. The experimental outcomes indicate that the system can successfully detect weapons, track intruders, and generate timely reports for emergency response. These findings highlight the potential of integrating advanced AI techniques for enhancing home security, offering both practical and innovative solutions to address the challenges posed by property crimes in urban areas.

6. Discussion

To further enhance the proposed system, several improvements could be made across different components to increase accuracy, efficiency, and overall reliability in real-world scenarios.

Firstly, YOLO's object detection performance could be optimized by integrating a more advanced version of YOLO, such as YOLOv4 or YOLOv5, to improve detection

accuracy, particularly for smaller objects or in low- resolution images. Additionally, incorporating techniques like multi-scale detection or image augmentation could further improve object localization in challenging conditions, such as poor lighting or cluttered environments. Integrating faster and more precise object tracking algorithms, like SORT (Simple Online and Realtime Tracking), could help maintain identification continuity in real-time surveillance.

Secondly, to enhance the RNN-family models (RNN, LSTM, GRU), it would be beneficial to incorporate transformer-based models, which have shown superior performance in processing sequential data. Transformers, with their attention mechanisms, allow for better handling of long-range dependencies and can improve the system's ability to understand complex, context-dependent messages. This could significantly improve the filtering and interpretation of emergency messages, particularly in noisy or ambiguous situations.

Additionally, integrating sensor fusion techniques, where data from different sources (e.g., motion sensors, temperature sensors, doorbell cameras) are combined, could provide a more holistic and robust detection mechanism. Using multimodal data (combining visual, auditory, and motion data) would allow for more accurate threat identification and reduce the occurrence of false alarms.

Finally, enhancing the real-time response system by incorporating an automatic emergency alert mechanism, such as a direct integration with local police or neighborhood watch systems, could speed up the process of emergency intervention. A predictive analytics model could be added to the system, leveraging historical data to anticipate potential security risks based on patterns of intrusion or past criminal behavior in the area, providing proactive security measures.

By implementing these improvements, the system's accuracy, response time, and overall robustness in preventing and mitigating home invasion crimes could be significantly enhanced, making it an even more reliable tool for modern home security.

References

[1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779-788. https://doi.org/10.1109/CVPR.2016.91

[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short- term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[3] Cho, K., van Merriënboer, B., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase

representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1724-1734.

[4] Zhang, Y., Zhao, Y., & Guo, Y. (2019). Real-time object detection with YOLOv3. Journal of Computer Science and Technology, 34(3), 633-642. https://doi.org/10.1007/ s11390-019-1941-x

[5] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. Proceedings of the 2014 NIPS Workshop on Deep Learning and Representation Learning.

[6] Li, X., & Lin, J. (2018). RNN for sequence learning and time-series prediction. Journal of Artificial

Intelligence Research, 59(1), 1-22. https://doi.org/10.1613/jair.5720
[7] Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), 1440-1448. https://doi.org/10.1109/ICCV.2015.169
[8] Wu, Y., & He, K. (2016). Group normalization. Proceedings of the 2018 European Conference on Computer Vision (ECCV), 3-19.
[9] Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial examples. Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 32-46. https://doi.org/10.1109/ EuroSP.2016.29
[10] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015).

Seowoo Choi
He is currently a junior attending Seoul International School, with research interests in Criminology, data mining, and social sciences.