EECE 570 Project Report
Enhancing Digital Privacy: Utilizing YOLOv8n for Sensitive Information Detection in WeChat Screenshots

Sara Zhang
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, BC Canada V6T 1Z4
[redacted]
zhangxiyu100@gmail.com

Abstract

This report details the development of a deep learning model to detect personal information in WeChat screenshots. Leveraging the YOLOv8n object detection framework, the model addresses privacy concerns associated with sharing digital images. This work not only explores technical aspects of adapting YOLOv8n for a specialized task but also evaluates its performance and discusses potential enhancements for real-world applications.

1 Introduction

The use of digital communications carries the risk of inadvertent disclosure of personal information, particularly in the sharing of images such as screenshots. This project employs the YOLOv8n model to automatically detect sensitive elements in WeChat screenshots which makes it possible to anonymize them to prevent privacy breaches. The objective is to integrate deep learning techniques with effective detection capabilities to enhance user privacy in digital media sharing.

2 Related Work

2.1 Advancements in Object Detection Technologies

Object detection technology has made significant strides with the development of deep learning architectures designed for real-time processing. One notable innovation is You Only Look Once (YOLO), which revolutionized object detection by framing it as a regression problem that predicts bounding boxes and class probabilities directly from full images in one evaluation Redmon et al. [2016]. Following this, the introduction of Faster R-CNN improved detection speed through a Region Proposal Network that shares convolutional features with the detection network, significantly enhancing the efficiency of the detection process Ren et al. [2015] . Further refining this landscape, MobileNets was optimized for mobile and embedded applications, utilizing depth-wise separable convolutions to build lightweight yet powerful neural networks Howard et al. [2017].

The field of object detection has witnessed transformative changes with the advent of deep learning. Research demonstrates the adaptability of YOLOv8n in specialized environments like underwater target detection, showcasing the model's broad applicability Liu et al. [2023] . Similarly, YOLOv8n has been adapted to monitor cognitive engagement in educational settings, underscoring the model's versatility across different fields Xu et al. [2023].

Preprint. Under review.

2.2 Harnessing Object Detection for Social Media Privacy

Digital records, including text messages, videos, and social media posts, have a profound impact on everything from casual gossip to legal criminal investigations. One prevalent method for distributing these records is through the use of screenshots Shore [2023]. However, at the same time, there is an emergent, inductive potential of data leaks, affirming the critical need for robust privacy measures in digital communications Micali [2018].

Applications of Object Detection, especially in contexts focused on privacy, include the anonymization of faces in public imagery and the redaction of personal data in automatically processed documents. Moreover, the integration of object detection models such as YOLOv8n into the realm of social media privacy represents a burgeoning area of focus. This project introduces a novel application by employing deep learning model to detect and obscure sensitive content, thereby enhancing privacy for users on sharing screenshots of WeChat. The potential implications of this approach are substantial, potentially establishing new standards for the use of advanced machine learning techniques in maintaining digital privacy across diverse platforms.

3 Methodologies

3.1 Model Overview

The model training was conducted using the YOLOv8n variant, optimized for environments with constrained computational resources. This training involved 80 epochs on a pre-trained model obtained from Ultralytics, utilizing Google Colab for computational support. The training process was designed to automatically save results and performance metrics, facilitating ongoing monitoring and evaluation of the model's effectiveness.

The choice of YOLOv8n was strategic, leveraging its proven efficacy in real-time object detection with state-of-the-art architectures that enhance feature extraction and object recognition accuracy. Its anchor-free split Ultralytics head is particularly notable for enhancing detection accuracy, which is crucial for the precise identification of sensitive information in WeChat screenshots. This feature simplifies the detection process, making it highly suitable for applications demanding rapid processing and high accuracy.

3.2 Loss function

In the YOLOv8 model, the overall performance is quantified by three key loss components, each reflecting a specific aspect of the object detection process. The Box Loss, with a weight of 7.5, is critical for measuring the precision with which the model predicts the coordinates of bounding boxes, emphasizing the accuracy of object localization. The Classification Loss, assigned a lower weight of 0.5, is crucial for ensuring that each detected object is correctly identified, highlighting the model's accuracy in class recognition. Additionally, the Distribution Focal Loss (DFL), weighted at 1.5, addresses class imbalance by focusing on fine-grained classification tasks, enhancing the model's sensitivity to less frequent classes. Together, these losses not only individualize the evaluation of different detection aspects but also collectively contribute to the model's total loss, thereby providing a comprehensive measure of the model's overall effectiveness in object detection. This integrated approach to loss calculation enables a balanced optimization of localization, classification, and handling class imbalance.

3.3 Evaluation Metrics

In this project, various performance metrics were employed to evaluate the YOLOv8n model's capability in accurately detecting personal information in WeChat screenshots. Precision (P) measures the accuracy of detected objects, indicating the proportion of correct detections among all detections made, which is crucial for ensuring that the anonymization process does not erroneously obscure irrelevant information. Recall (R) assesses the model's ability to identify all relevant instances of objects within the images, a critical factor for comprehensive privacy protection as it reflects the system's sensitivity to potential privacy breaches. Mean Average Precision at 50% IoU (mAP50) focuses on the model's accuracy in relatively straightforward detection scenarios, providing insight into its efficiency under less challenging conditions. Furthermore, Mean Average Precision from

50% to 95% IoU (mAP50-95) averages precision across a range of stricter IoU thresholds, offering a detailed view of the model's performance across varying levels of detection difficulty, from easy to highly challenging scenarios. Lastly, the F1 Score serves as a harmonic mean of precision and recall, presenting a balanced measure of the model's overall test accuracy. This comprehensive set of metrics is vital for evaluate the effectiveness of the model in detecting sensitive content and in ensuring that such detections are accurate and reliable.

4 Experiments

4.1 Dataset

The dataset utilized for this project consisted of 130 manually annotated WeChat screenshots, containing a total of 510 instances of personal information. The images were divided into training, validation, and testing subsets, with 93, 25, and 12 images respectively. These screenshots, collected over a span of five years, encompass a diverse array of conversations, including both individual and group chats. The variability in the dataset is further enhanced by the inclusion of different user interfaces over the years, which feature varying backgrounds, profile photos, and layouts of user names. For example, group chats often display multiple user profiles and names above each message depending on the user settings, while one-to-one chats typically show fewer profile icons and may only display user names once.

Due to the sensitive nature of the content, existing public datasets for chat screenshots, especially those specific to WeChat, are scarce. This scarcity, primarily driven by privacy concerns, necessitated the creation of a custom dataset. The annotations were meticulously performed using the Computer Vision Annotation Tool (CVAT), a process that involved identifying and labeling sensitive personal information such as usernames and profile photos. Figure 5 shows an example of annotated image. This manual annotation process was both time-consuming and required a high degree of precision to ensure the accuracy of the bounding boxes around sensitive information.

This project employs the default configuration settings provided by YOLO for data preprocessing and augmentation, capitalizing on the straightforward and relatively clean nature of the collected data source. Data augmentation is fundamental in boosting the robustness and performance of the YOLOv8n model, injecting variability into the training set which aids the model in adapting to new, unseen scenarios. This process involves modifying hue, saturation, and brightness to reflect different lighting conditions and applying geometric transformations such as rotation, translation, scaling, and shearing, which enhance the model's capability to recognize objects from various perspectives and orientations. Additional variability is introduced through random flips and channel switching. Sophisticated techniques like mosaic and mixup intermingle images and labels, which enriches scene comprehension and label precision. Techniques such as copy-paste and erasing are employed to train the model to detect objects from incomplete views and concentrate on essential features, respectively. Auto augment policies and adjustable cropping techniques are also utilized to ensure the model prioritizes key object features while reducing background noise.

4.2 Model Training

The training regime employed for enhancing the detection of sensitive information was substantially anchored in the default configurations as recommended in the YOLO documentation.

The batch size of 16 was selected as it is both computationally manageable and effective in updating model parameters. Image inputs of 640x640 pixels were normalized, representing a fine balance between computational complexity and detection accuracy. A learning rate of 0.01 was employed, with this value gradually decreasing to 1% at the conclusion of the training process via the use of a learning rate scheduler. This scheduler was designed to gradually smooth the model towards the minimization of loss. Additionally, a momentum of 0.937 was utilized to assist in stabilizing convergence by combining prior gradient updates with the current one. To mitigate the risk of overfitting, L2 regularization, which aims to regulate the magnitude of weights, was employed at a value of 0.0005. A warmup of three epochs trained the model such that the momentum was recouped from 0.8 to the specified value and the learning rate for bias parameters was reduced to 0.1, facilitating a gradual descent into training.

Validation was enabled (val=True) during the training. This enabled the assessment of the suitability of the colloquial use of models to a different validation dataset, with a view to monitoring the progress of the model and its predictive ability.

Figure 1: Losses during training process.

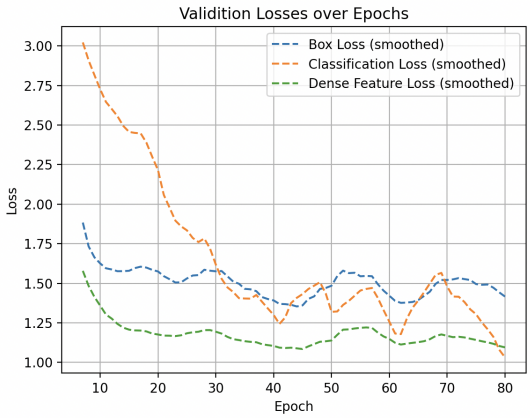

Figure 2: Losses during validation process.

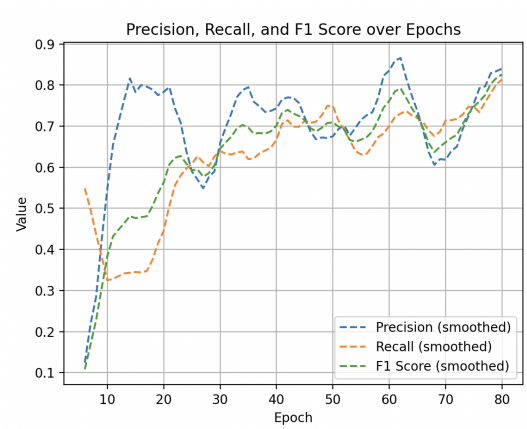

Figure 3: Precision, recall and F1 score during training process.

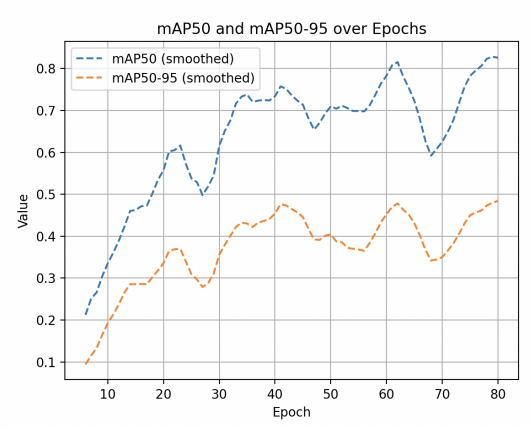

Figure 4: Mean Average Precision at 50% IoU (mAP50) (mAP50) and Mean Average Precision across IoU from 50% to 95% (mAP50-95) during training process.

As illustrated in Figure 1, the training of the YOLOv8n model resulted in high losses initially. However, these losses rapidly declined, indicating effective learning and generalization across various characteristics of WeChat data. These outcomes were consistent between the testing and validation datasets, suggesting that our model can generalize broadly without overfitting.

Figure 3 shows that the precision, recall, and F1 score stabilized around 0.75, suggesting the model accurately identifies and minimizes false positives, a crucial factor for privacy applications where incorrectly labeling data as sensitive could lead to excessive censorship. However, the distinction in mAP scores between simpler scenarios (mAP50) and more complex conditions (mAP50-95) highlighted in Figure 4 indicates the model's robust performance in clear cases but reveals struggles in more ambiguous contexts, such as overlapping texts or multimedia elements.

These performance measures necessitate further optimization in the complexity of interactions commonly seen within WeChat. Such strategies may include improving the training dataset with more challenging examples in larger datasets, utilizing a more robust contextual analysis toolkit for increased accuracy and application reliability.

4.3 Performance Evaluation

After training, the YOLOv8n model was evaluated on a test set of 12 images, containing 47 instances. The model achieved a precision of 96.8%, indicating a high accuracy in correctly identifying instances when predicted. The recall was 85.1%, reflecting the model's ability to detect most relevant instances. Mean Average Precision at a 50% Intersection over Union (IoU) threshold (mAP50) was impressively high at 95.2%, demonstrating accurate bounding box predictions. However, the mAP across IoU thresholds ranging from 50% to 95% (mAP50-95) was 65.8%, showing a performance decline at stricter IoU thresholds. These metrics suggest a robust model performance at typical settings but indicate room for improvement under more stringent conditions.

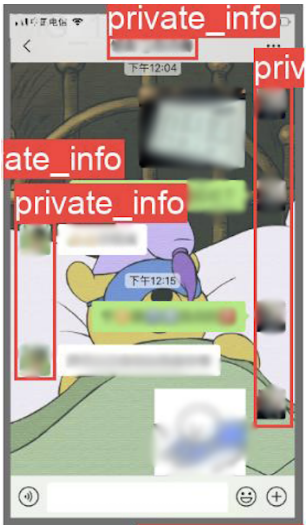

Figure 5: Annotated Test Screenshot: Display of the original WeChat screenshot annotated with ground truth bounding boxes for sensitive information. Note that the image is blurred before displayed for privacy concerns.

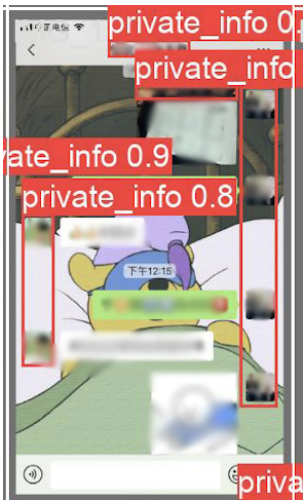

Figure 6: Model Detection Results: Output from the trained model showing detected sensitive information within the WeChat screenshot, highlighted with bounding boxes. Note that the image is blurred before displayed for privacy concerns.

In Figure 5 and 6, we illustrate the comparison between ground truth annotations and model outputs for a sample test screenshot. The left image shows the test screenshot with annotations marked as ground truth, which serves as a reference for evaluating model accuracy. The right image presents the detection results obtained from the final model, highlighting the model's performance in accurately identifying and localizing objects.

The provided confusion matrix in Figure 7 illustrates the performance of the classification model on the test set, with a focus on discerning 'private_info' from 'background'. The matrix indicates a robust ability to correctly identify 'private_info', evidenced by 43 true positives, though it also shows a tendency to misclassify 'background' as 'private_info' in 15 cases, suggesting a potential oversensitivity to the 'private_info' class. Moreover, there are 4 instances where 'private_info' is wrongly classified as 'background', hinting at occasional under-detection. While the model demonstrates high precision in identifying private information, the presence of both false positives and negatives highlights the need for further refinement, particularly to reduce over-censoring and enhance the model's recall.

5 Conclusion, Challenges and Future Directions

This project demonstrates the potential of using advanced object detection models like YOLOv8n to enhance privacy in digital communications. By effectively detecting personal information in WeChat screenshots, the model provides a viable solution to a significant privacy issue.

Among the challenges faced were the manual effort required for accurate data annotation and the adaptation of the YOLO model to handle densely packed text objects. The project's scope is currently

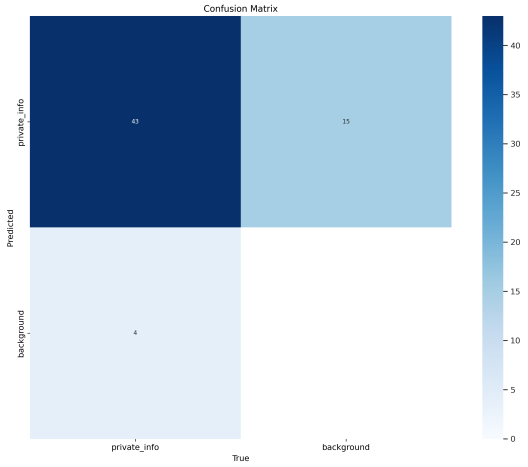

Figure 7: Confusion Matrix displaying the true positive and false positive detections of 'private_info' versus 'background' by the trained YOLOv8n model.

limited by the dataset size and variability. Future improvements could involve expanding the dataset, incorporating more diverse data scenarios, and implementing the model in a real-time system for automatic processing of images before they are shared.

Ensuring annotation consistency across the dataset posed significant challenges, impacting the model's learning efficiency and accuracy. The variability in how different elements were labeled led to inconsistencies in training, affecting the model's ability to generalize to new, unseen data. Additionally, due to the sensitive nature of the data, there were inherent privacy concerns that limited the diversity and breadth of the data available for training. This limitation was especially critical during online training sessions, where data security and privacy considerations are paramount.

Looking ahead, future work will focus on exploring diverse neural network architectures and models to enhance the detection capabilities and improve the system's robustness against various types of inputs and environmental conditions. Expanding the dataset to include screenshots from other applications and different devices is crucial. This diversity will enable the model to better recognize personal information across a broader spectrum of social media interfaces and technological platforms. Additionally, the further step in future project involves integrating the trained model into a complete pipeline. This pipeline will not only detect but also blur or block identified sensitive information, facilitating real-time privacy protection and making the solution practical for everyday use by individuals and organizations.

References

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017.

Qiang Liu, Wei Huang, Xiaoqiu Duan, Jianghao Wei, Tao Hu, Jie Yu, and Jiahuan Huang. Dsw-yolov8n: A new underwater target detection algorithm based on improved yolov8n. *Electronics*, 12(18), 2023. ISSN 2079-9292. doi: 10.3390/electronics12183892.

Alberto Micali. Leak early, leak (more than) often: Outlining the affective politics of data leaks in network ecologies. *Media and Communication*, 6(3):48–59, 2018. ISSN 2183-2439. doi: 10.17645/mac.v6i3.1440.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

Prena K. Shore, A. Platform rules as privacy tools: The influence of screenshot accountability and trust on privacy management. *New Media Society*, 215, 2023. doi: 10.1177/14614448231188929.

Qi Xu, Yantao Wei, Jie Gao, Huang Yao, and Qingtang Liu. Icapd framework and simam-yolov8n for student cognitive engagement detection in classroom. *IEEE Access*, PP:1–1, 01 2023. doi: 10.1109/ACCESS.2023.3337435.