C951 Task 3 Writeup:
Machine Learning Project Proposal

Xavier Loera Flores

ID:011037676

xloeraf@wgu.edu

C951 Introduction to Artificial Intelligence

C951 Task 3 Writeup:
Machine Learning Project Proposal....1

Project Overview

The following project proposal will demonstrate the need for a machine-learning solution that implements image recognition to address an organizational need. The proposal will include a project overview, project design, machine learning solution design, and a description of the data sets that will be used to train the machine learning model.

Organizational Need

A social media company needs a machine learning solution to automatically detect what type of content is in an image to flag the post as explicit or not explicit for moderation and content filtering purposes. The platform needs to be able to automatically detect the content of images as the user uploads the images to the platform.

Context and Background

The company has seen a significant increase in the number of images with explicit content being uploaded to their platform. While the platform allows users to manually mark their posts as NSFW or explicit, many users are not doing so, and the company receives numerous complaints from users who do not wish to see explicit content. This will allow the moderation team to focus on images with explicit content and remove them from the platform. It would also allow the company to blur or hide explicit images from users who do not wish to see them.

Outside Works

The following works were used to help guide the proposal and provide context to the project.

Work 1

"You Only Look Once: Unified, Real-Time Object Detection" by Joesph Redmon, et al. (1):

This paper was used to understand the YOLO algorithm and how it can be used to detect objects in images. The YOLO algorithm is a popular algorithm for object detection and can be used to detect explicit content in images. The paper highlights that the YOLO algorithm is capable of processing images in real time at 45 frames per second. It features a smaller version of the network called Fast YOLO that processes images at a rate of 155 frames per second.

Work 2

"Rich feature hierarchies for accurate object detection and semantic segmentation" by Ross Girshick, et al. (2):

This paper as presented by UC Berkeley, highlights the R-CNN algorithm and its utility when it comes to object detection in images. The authors describe R-CNN as regions with CNN features since the method combines region proposals with convolutional neural networks to detect objects in images. The paper also compares the R-CNN algorithm to another algorithm called OverFeat which also uses a CNN architecture but utilizes a sliding window approach to detect objects in images. The authors highlight how R-CNN outperforms OverFeat (31.4% vs 24.3% respectively) in terms of mean average precision (mAP) meaning it is able to detect objects in images with a higher degree of accuracy.

Work 3

"An Algorithm for Nudity Detection" by Rigan Ap-apid (3):

This paper was used to understand the process of detecting nudity in images. It highlights the use of detecting skin in images across different color formats. Using correlation and linear regression, a skin color distribution model is created to identify skin regions in images. The regions are then analyzed for clues that indicate nudity such as the presence of a face, body, or other body parts.

Machine Learning Solution

While a solution such as the Nudity Detection algorithm by Rigan Ap-apid (3) could be used to detect nudity in images, the company needs a solution that can detect all types of explicit content such as violence, hate speech, and other types of explicit content. A nudity detection approach is too specialized for the purposes of our machine learning solution since we not only need to detect nudity but also violence and other explicit material that does not fall under the category of nudity. The company also needs a solution that can process at least 100,000,000 images per day and cost less than $1.5M per month to operate. The company also needs a solution that does not negatively impact the user experience for uploading images. The plan is to choose a fast object detection algorithm that can be trained to not only detect nudity but other explicit material as well. The machine learning solution should be run to automatically flag images as SFW or NSFW. Therefore, the company will be using a machine learning solution more in generalized object detection such as the R-CNN or the YOLO algorithm. The following section highlights the benefits of using our chosen solution, the YOLO algorithm over the R-CNN algorithm.

Benefits of Proposed Machine Learning Solution

The YOLO algorithm paper showcases benchmarks that are in favor of the YOLO algorithm in comparison to an R-CNN algorithm. The paper highlights that since the YOLO algorithm looks at the entirety of an image all at once rather than looking at regions or a sliding window, the YOLO algorithm makes less than half the amount of background errors compared to the Fast R-CNN algorithm. As stated before, the YOLO algorithm can process in real-time at up to 45 frames a second and at 155 fps on a faster version of the network which means YOLO is capable of real-time processing while R-CNN is not. The benefit of using the YOLO algorithm in our machine learning solution that will be run with every image upload is that the YOLO algorithm is fast enough to handle

real-time processing with only about 25 milliseconds of latency. This means that users should be able to upload images without having to wait long for algorithms like R-CNN to take up to 2 seconds to process an image.

Machine Learning Project Design

Scope

The scope of this project is to develop a machine-learning solution that can accurately and automatically detect explicit content in images uploaded to the company's social media platform. The scope includes the following:

- Collecting and preparing the data
- Training the machine learning model
- Integrating the model with the platform
- Testing the model
- Deploying the model
- Monitoring the model The scope does not include the following:
- Developing a new interface for the moderation team
- Retraining the model with new data
- Detecting and categorizing the types of explicit content such as nudity, violence, or hate speech
- Detecting explicit content in videos
- Detecting explicit content in text on the images

Goals, Objectives Deliverables

The goal of this project is to develop a machine-learning solution that can detect explicit content in images uploaded to the company's platform. It should be able to achieve the following objectives:

Accurately categorize images as NSFW or SFW:

The machine learning solution needs to be able to accurately categorize images as NSFW or SFW with a degree of accuracy that is beneficial and acceptable to the company's moderation team. If the model is not at least 95% accurate, it will not be useful to the moderation team, will not be implemented, and may even be a hindrance to the operations of the company's social media platform.

Process at least 100,000,000 images per day:

The machine learning solution needs to be able to process at least 100,000,000 images per day to be able to keep up with the throughput of images that are uploaded to the company's platform. If the model cannot keep up with the throughput of images, it will not be able to be implemented into the platform nor keep up with any future growth in the platform. Cost less than $1.5M per month to operate: The machine learning solution needs to cost less than $1.5M per month to operate to be able to be implemented and maintained by the company. If the model costs more than $1.5M per month to operate, it may not be approved by the company leaders as the solution does not help drive net revenue for the company.

Not negatively impact the user experience for uploading images:

The machine learning solution should not negatively impact the user experience for uploading images. If the model negatively impacts the user experience, it will not be implemented into the platform as it will drive users away from the platform.

Standard Methodology

Cross-Industry Standard Process for Data Mining, also known as (CRISP-DM) (4), will be used as the standard methodology for the project since it is a widely used and accepted methodology for machine learning projects and provides a structured approach for proceeding with the project. CRIPS-DM is a 6-phase process that includes the following phases with a description of how they will be used in the project:

- Business understanding: It is important to understand the specific needs of the company and the user's needs for the machine learning solution in relation to the company's platform. This will help guide the project and ensure that the machine-learning solution meets the needs of the company and its users.
- Data understanding: We will need to have a deep understanding of the types of images that will be used to train the machine learning model. It is best to use images that are representative of the images that are uploaded to the company's platform to ensure that the model is able to accurately categorize the images that are uploaded to the platform.
- Data preparation: The data will need to be prepared for the training of the machine learning model through cleaning, formatting, and anomaly mitigation. The images will be cleaned to remove any irrelevant, spam, duplicate, or miscategorized images.
- Modeling: The machine learning model will be trained and optimized to accurately categorize images as NSFW or SFW. The model will be trained to accept user-generated content by being trained on user-generated content.
- Evaluation: The machine learning solution will be evaluated to ensure that it meets the goals and objectives of the project. The model will be deployed to a small subset of users to monitor the model and collect feedback which is needed to evaluate the model since our goal is to not negatively impact the user experience for uploading images.
- Deployment: The machine learning solution will be deployed to the entire platform and monitored to ensure that it continues to meet the goals and objectives of the project.

Project Timeline

A rough estimate of the project timeline is as follows:

Sprint Timeline Overview:

Sprint	Dates	Task

1	04/01/2024 - 04/08/2024	Project Planning
2	04/08/2024 - 04/22/2024	Data Collection
3	04/22/2024 - 05/06/2024	Data Preparation
4	05/06/2024 - 05/27/2024	Model Training
5	05/27/2024 - 06/10/2024	Integration & Testing
6	06/10/2024 - 06/17/2024	Evaluation
7	06/17/2024 - 07/01/2024	Model Deployment

Sprint 1: Project Planning (1 Week)

- Define the project scope and goals
- Define the project timeline
- Define the project budget
- Define the project team

Sprint 2: Data Collection (2 Weeks)

- Identify the data sources
- Write the data collection method
- Collect the data
- Store the data

Sprint 3: Data Preparation (2 Weeks)

- Clean the data
- Format the data
- Mitigate data anomalies
- Validate the data

Sprint 4: Model Training (3 Weeks)

- Set up the machine learning environment
- Train the machine learning model
- Debug the model

- Optimize the model
- Validate the model

Sprint 5: Integration & Testing (2 Weeks)

- Integrate the model with the platform
- Write tests for the model
- Test the model
- End-to-end testing
- Create a maintenance plan

Sprint 6: Evaluation (1 Week)

- Deploy the model to a small subset of users
- Monitor the model
- Collect feedback
- Evaluate the model

Sprint 7: Model Deployment (2 Weeks)

- Deploy the model to the entire platform
- Monitor the model
- Collect feedback
- Document the model

Resources and Costs

Using estimates from Microsoft Azure Visions Pricing (5) as well as the daily estimates for the number of images uploaded to Instagram from Bernard Marr (6), we can estimate the cost of the machine-learning solution.

Assuming:

- 100,000,000 Daily uploaded images
- $0.40 Per 1000 Images at a rate of at least 1,000,000 transactions per day

Resource	**Cost**
Engineering Labor	$100K-150K
Database Storage	$10,000
Data Processing	$10,000
Machine Learning Model Training	$10,000

Upfront Cost Total	~$130K-180K
Machine Learning Cloud Server	~$35K-45K daily
Maintenance Engineering Labor	$10K-15K monthly
On-Going Total	~$1M - 1.35M monthly

Success Criteria

Objective	Success Criteria
Model Accuracy	The machine learning solution should accurately categorize images as NSFW or SFW with at least 95% accuracy.
Model Performance	The machine learning solution should be able to process at least 100,000,000 images per day.
Model Cost	The machine learning solution should cost less than $1.5M per month to operate.
User Experience	The machine learning solution should not negatively impact the user experience for uploading images within a 5% margin of error feedback score.

Machine Learning Solution Design

Proposed Project Hypothesis

The company needs a machine learning solution that can automatically detect explicit content in images uploaded to the company's social media platform. The company can utilize a machine learning solution that implements the YOLO algorithm, a fast object detection algorithm, to automatically detect whether images contain explicit material. The machine learning solution will allow the company to automatically flag images as NSFW or SFW for moderation and content filtering purposes with great accuracy, throughput, and cost efficiency without negatively impacting the user experience for uploading images.

Machine Learning Algorithms

The chosen machine learning approach will be to use supervised learning with the You Only Look Once or YOLO algorithm. The YOLO algorithm provides a good balance of speed and accuracy while being robust enough to not only handle detecting nudity but also detect other types of explicit material as well regardless if they are in the background or in the foreground. We will be using supervised learning because we already have a labeled and categorized dataset of images to train the machine learning model on.

Algorithm Advantage

The YOLO algorithm's main advantage for our context is its speed and capability to process images in real-time. With only about 25 milliseconds of latency, our users will be able to upload images without having to wait long for the algorithm to process the image. The YOLO algorithm is also better at detecting images in the background resulting in less than half the error rate of R-CNN.

Algorithm Limitation

While the YOLO algorithm is a very capable algorithm when it comes to speed and objects in the background, it still falls short of the Fast R-CNN algorithm's accuracy with objects in the foreground with a mAP of 63.4 vs the Fast R-CNN mAP of 71.8. The YOLO algorithm should also in a nudity detection context fall short of a specialized nudity detection algorithm such as the one presented by Rigan Ap-apid (3) since the YOLO algorithm is not specialized for nudity detection.

Tools and Environments

The following tools and environments will be used to develop the machine-learning solution:

- Python: The machine learning model will be developed using Python since it has the most support for machine learning libraries and frameworks.

- TensorFlow: TensorFlow will be used to develop the machine learning model since it is a widely used and supported machine learning framework.
- Linux: The machine learning model will be deployed on a Linux server since it is a widely used and supported operating system for machine learning models.
- Microsoft Azure server: Microsoft Azure will be used to host the machine learning model since the company already hosts their platform on Microsoft Azure and it is a widely used and supported cloud platform for machine learning models.
- Microsoft Azure Storage: Microsoft Azure Storage will be used to store the image data that will be used to train the machine learning model.

Performance Measuring Process

The machine learning solution needs to be measured in regard to speed, accuracy, and cost. Using a simulated scaled environment, the machine learning solution will be tested to ensure that it meets the goals and objectives of the project before it is deployed and scaled to the entire platform. It needs to handle a throughput of at least 100,000,000 images per day, cost less than $1.5M per month to operate, maintain an accuracy of at least 95%, and not negatively impact the user experience for uploading images. While accuracy, speed, and cost can be tested in a simulated environment, the user experience will need to be tested with a small subset of users before the model is deployed to the entire platform.

Description of Data Sets

Data Sources

Data will be sourced from all the public images uploaded to the company's social media platform by users. The company's terms of service allow for the use of images uploaded to the platform for improving features on the platform. These images can be found on the company's database and can be accessed through the company's API or either through a direct connection to the database.

Data Collection Method

Once we have access to the collection of image data via either the company's API or direct connection to the database, we will collect the images to store them for use in training the machine learning model. For the most part, the data collected will already be categorized as explicit or not explicit since the company's moderation team had already flagged the images when the throughput of images was lower. The company only recently faced the issue of an increase in un-flagged explicit images so most of the images on the platform are already categorized.

Method Advantage

By utilizing the database of images from the company's platform, we ensure that the data that will train the machine learning model is representative of the data that the model will be used on. The model will be trained to accept user-generated content by being trained on user-generated content. This will allow the model to tackle edge cases that can be found in user-generated content since the model will be trained on years worth of data from the company's platform.

Method Limitation

The limitation of this method is that the model will be trained on data that has already been flagged by the company as explicit. User-generated content can be unpredictable and new types of explicit content can be uploaded that the platform will not recognize such as AI-generated images, deep fakes, and hand-drawn explicit images. The model will be able to detect previous trends in explicit content but may not be able to detect new types or even subtle instances of explicit content.

Data Preparation

The data will need to be prepared for the training of the machine learning model through cleaning, formatting, and anomaly mitigation. The data will be cleaned to remove any irrelevant, spam, duplicate, or miscategorized images. All the images will be formatted to the standard canvas and file size so that images are stored and viewed on the platform. Using these goals in mind, checks and validation will be run on the data to mitigate data anomalies and ensure that the data is consistent and ready for training.

Behaviors for Handling & Communicating Data

It is important that when handling the data we are transparent with the company and the users about the use of their data. This needs to be done with respect for the users because some users may not want their images to be used for training machine learning models. Once the model is implemented, the images should be user anonymized and the model should not store any images that are being processed in the future. This is because the model will be used on images that are uploaded to the platform which some users may eventually want to delete. We may eventually revisit and retrain the model with new data later on to improve the model's accuracy and performance.

References

(1) Redmon Joesph, et al. "You Only Look Once: Unified, Real-Time Object Detection" University of Washington https://arxiv.org/pdf/1506.02640.pdf Accessed Feb 16,2024.

(2) Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation" University of California Berkeley https://arxiv.org/pdf/1311.2524v5.pdf Accessed Feb 16,2024.

(3) Ap-apid, Rigan "An Algorithm for Nudity Detection" College of Computer Studies De La Salle University https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6d52a6fc245a22b35cb6373ff9ce29a033b57a5b Accessed Feb 16,2024.

(4) Hotz, Nick "What is CRISP DM?" Data Science Process Alliance https://www.datascience-pm.com/crisp-dm-2/ Accessed Feb 20,2024.

(5) Microsoft. "Microsoft Azure Pricing" Microsoft Azure https://azure.microsoft.com/en-us/pricing/details/cognitive-services/computer-vision/ Accessed Feb 24,2024.

(6) Bernard Marr & Co. "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read" BernardMarr.com https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/ Accessed Feb 24,2024.