Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

R&D Project Proposal

Multi-input Object Detection with Encoding Labels and Fusion Techniques using Conditional Generative Adversarial Networks

Gokul Krishna Gandhi Chenchani

Supervised by

Prof. Dr.-Ing. Sebastian Houben
M.Sc. Deebul Sivarajan Nair

September 2023

1 Introduction

Deep learning has transformed disciplines such as computer vision, natural language processing, and speech recognition. Deep learning's capacity to handle the representations of complicated data is one of its main features, making it appropriate for problems involving numerous data modalities.

Currently there are numerous deep learning approaches in use with different architectures for detection and classification of objects for a given input image. Each approach has its set of parameters based on which the model is trained for a given dataset. The common deep learning approaches for object detection using a trained model on a dataset has few challenges when the goal is to detect one object and the input image contains multiple objects which belong to different classes, the output is set to detect all the possible classes present in the given input image. The tensor computation also varies based on the architecture and the processing of input data through the trained model . There have been few studies done on multiple input deep learning approaches such as open vocabulary object detection[14], conditional deep learning[17] and conditional generative adversarial networks (cGAN's)[15].

Generative adversarial network (GAN)[9] trains on a generator and a discriminator where the generator model tries to counter-fit the data or generate fake data and the discriminator model is used to detect this fake data then trying to eliminate the possibility of wrong identification and detect a valid class from the inputs provided.

A conditional generative adversarial network (cGAN) is an approach based on the GAN with just an addition of a conditional input to the model. This additional input (in our case an encoded label) is used by the model to process the data (in our case an image with multiple classes) based on the condition.[7]

In this research and development project, I am going to implement the concepts inspired from conditional generative adversarial networks (cGAN's) by passing multiple inputs like image and encoded label to the model to generate an output that only detects the object of the input label passed among all the other trained classes.

The proposed model is trained on two datasets one being a primary dataset of

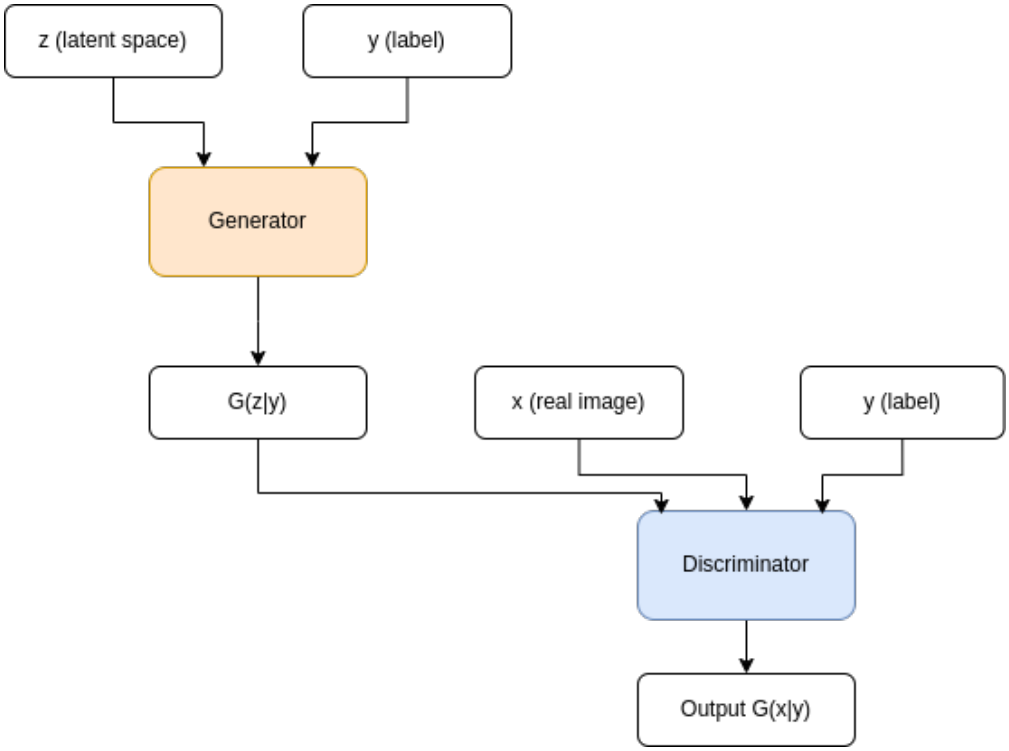

Figure 1: CGAN architecture with image and label input[11]

RoboCup@Work[6][13] (industrial objects) and the other being a secondary dataset of HOPE[21] (household objects with 6-DoF).

1.1 Relevance of This R&D Project

The primary objective of this R&D project is to focus on the object detection of RoboCup@Work tasks and try to reduce the computational power to detect the object more efficiently and faster than the current used model. The results obtained in this project will be tested thoroughly on the RoboCup@Work dataset and also by capturing images of the objects in real-time and evaluating the performance of the system by testing the implementation on YouBot[1] at b-it-bots@Work[12] using RealSense cameras. This implementation will also be tested on the secondary dataset of HOPE as mentioned above.

The benefits and study done for this project will be helpful in implementing

[1] http://www.youbot-store.com/developers/kuka-youbot-kinematics-dynamics-and-3d-model

the proposed approach in YouBot at b-it-bots@Work for the RoboCup@Work 2024 competition by improving the system performance and efficiency of the object detection model to accurately identify and then manipulated the object based on the object coordinates and pose from the model. Also, one of the main challenges to be addressed using this approach is to determine the accuracy of detection of objects on arbitrary surfaces which has always been a challenge due to multiple factors during the competition.

2 Related Work

J Gauthier[8] discusses the development of the conditional generative adversarial network (CGAN) on a face image dataset used for convolutional face generation. This development works on generative adversarial network (GAN) with an arbitrary conditioning y, which is an embedding space. The architectural framework of this approach is shown in figure 2.

Mathematical representation of GAN is shown in equation 1 and the representation of CGAN is shown in equation 2

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z\sim p_z(z)} [\log(1 - D(G(z))] \tag{1}$$

$$\min_G \max_D \mathbb{E}_{x,y\sim p_{\text{data}}(x,y)} [\log D(x,y)] + \mathbb{E}_{y\sim p_y, z\sim p_z(z)} [\log(1 - D(G(z,y),y))] \tag{2}$$

Ezeme et al.[5] studies the design and implementation of CGAN for anomaly detection in non-parametric multivariate data. This work also studies the realistic distributions of a given dataset and solving the issue of imbalance in data in anomaly detection tasks. The study uses a single class CGAN and entails the process of learning the pattern of the minority class data samples. This knowledge can be used to detect the minority class samples, allowing a binary class CGAN to train with a balanced dataset on both normal and harmful characteristics. From the results observed in this study, AD-CGAN performs better than most algorithms in conventional measures such as Precision, Recall, and F-1 Score.

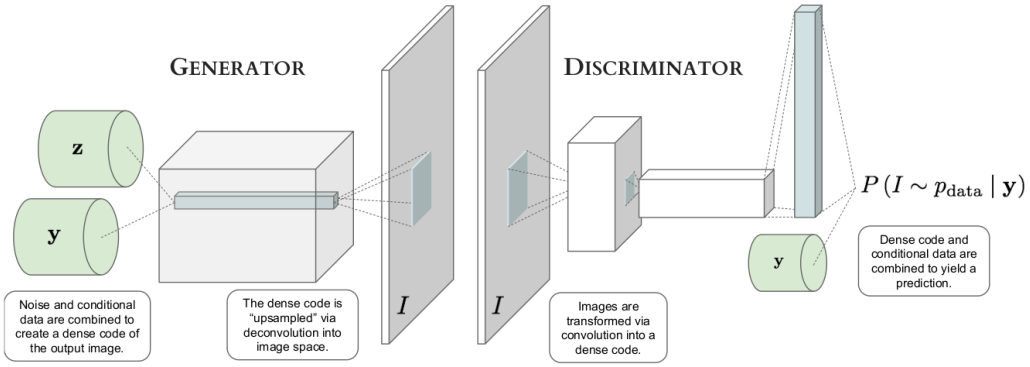

Figure 2: Graphical overview of the conditional generative adversarial network (CGAN) framework[8]

Hyojin Park et al.[18] studies another novel approach of using multiple conditions to detect an object using CGAN. In this approach, the model generates a new image by analyzing the backdrop of an original base picture and a new object defined by the text description in a specific location. The approach is implemented by a series of synthesis blocks in which the inputs constitute the seed feature map from a fully connected layers. Then the combination of image features from the backdrop image generates an output image and a segmentation mask. This study was carried out on Caltech-200 bird[22] and Oxford-102 flower[16] datasets.

Several other studies of CGAN were carried out in various domains such as abnormal vibration of aero engine analysis by Yang, Lu[23], fuse sar and multi-spectral optical data for cloud removal from sentinel-2 images by Grohnfeldt et al.[10]

3 Problem Statement

The current object recognition in YouBot (RoboCup@Work) is implemented using YOLOv8[19][20] model, where the input image is fed to the model and the output received is the list of all the classes identified along with the bounding boxes in the given image on which the model is trained upon which the object that needs to be manipulated is queried from the list and the co-ordinate details are sent to the

manipulator to grasp the object.

As per the existing implementation, the system knows which object to manipulate from the planner even before passing the image as input to the model. This sometimes can increase the tensor computation while identifying all the possible classes in the input image.

Another challenge while using the current YOLOv8 model on the RoboCup tasks is that the identification of objects on arbitrary surfaces is not very much impressive and sometimes also identifies an object with an incorrect class.

The proposed approach is to use the concept of CGAN and implement a trained model for which there are multiple inputs such as image and encoded label(using an encoded approach) along with the use of some fusion methods such as concatenating the latent vectors of input image and input label to the model to expect an output of the image with the detection of only the object of the input label class.

3.1 Research & Development Questions

3.1.1 Research Questions

- Which encoding deep learning approach can be used for comparative evaluation?
- Which fusion methods can be implemented for comparative evaluation?
- Why is this approach necessary in the current @Work scenario?
- Determine a comparative run-time analysis with current YOLOv8 model used in the YouBot for performance and computation power along with test on arbitrary surfaces.

3.1.2 Development Questions

- Determine the performance and computation power by deploying the final study and implementation on the real YouBot hardware.

3.2 Encoding Approaches

3.2.1 One-Hot Encoding

One-hot encoding is a method of converting data of all the class representations of the trained model into one new binary variable by denoting an integer 1 to the class that needs to be detected and 0's to all the other classes [4][2] .

3.2.2 Error-Correcting Encoding

Error-correcting encoding is a technique for breaking down a multi-way classification problem into several binary classification tasks. This method assigns a unique n-bit vector to each label of class size m (where n > log2 m). Each bit vector is considered a unique coding for a label, forming a code matrix, denoted by C. Each row, C_i, and the value of the jth bit in this row, C_{ij}, represent specific codings.[3]

3.3 Fusion Approaches

3.3.1 Early Fusion

Early fusion is the process of concatenating features from many modalities at an early stage before they are transferred to the modeling process units.

- Summation of latent vectors of image and label text into a single vector.
- Text to visual features transformation for a CNN model[1][7].

The proposed approach for the problem statement described above will be performed on both primary and secondary datasets and the results will be compared with the existing YOLOv8 model used in YouBot at b-it-bots@Work.

4 Project Plan

4.1 Work Packages

The R&D project will contain the following work packages

WP1 **Literature study**

- Conduct a comprehensive literature study on CGAN approach and its various use cases.
- Conduct a comprehensive literature study on encoding approaches.
- Conduct a comprehensive literature study on fusion techniques for image and encoded label.

WP2 **Experimental setup and analysis**

- Collect annotated dataset of RoboCup@Work and HOPE.
- Train a GAN model for both the datasets.
- Perform basic object detection on the trained models.

WP3 **Implementation of encoding and fusion approaches**

- Implement and analyse the performance of one-hot encoding on the trained model.
- Implement and analyse the performance of error-correcting encoding approach on the trained model.
- Implement and analyse the performance of early fusion technique of image and encoded label.

WP4 **Mid term report**

- Documentation of detailed review on the study of implemented encoding and fusion techniques.
- Perform various experiments on the trained model.

WP6 **Real-time testing of implemented approaches**

- Integrate the implemented approach with the YouBot object detection module.
- Testing of the implemented approaches using real-time images from the YouBot.

- Observing the performance results on the trained model for regular images and real-time images.

WP7 **Evaluation of approach and comparison with existing approach**

- Evaluate the performance of the system using proposed approach and compare with existing YOLOv8 model.
- Evaluate the computational power of the system using proposed approach and compare with existing YOLOv8 model.

WP8 **Project Report**

- Documentation of conditional generative adversarial network (CGAN) approach with relevance to the proposed R&D project.
- Documentation of state-of-the-art of the use of CGAN techniques.
- Documentation of state-of-the-art of the use of encoding approaches in deep learning.
- Documentation of state-of-the-art of the use of fusion techniques for image and encoded label in deep learning.
- Documentation of the proposed approach and the methodology.
- Documentation of the use of encoding approaches and fusion techniques used in the project.
- Documentation of the results obtained using the proposed approach.
- Documentation of the comparative results of the encoding approaches and fusion techniques used in the project.
- Documentation of the comparative results of the implemented approach with the existing YOLOv8 model on YouBot.
- Draft R&D report explaining the findings of the research.
- Final R&D report explaining the findings of the research.

4.2 Milestones

M1 Literature review completed and best practice identified

M2 Implementation of the encoding and fusion techniques on the GAN model

M3 Mid term report

M4 Performance study and testing on real-time of the implemented approaches.

M5 Evaluation and comparison of the implemented approaches with existing approach.

M4 Report submission

4.3 Project Schedule

The timeline for the project can be seen in the figure 3

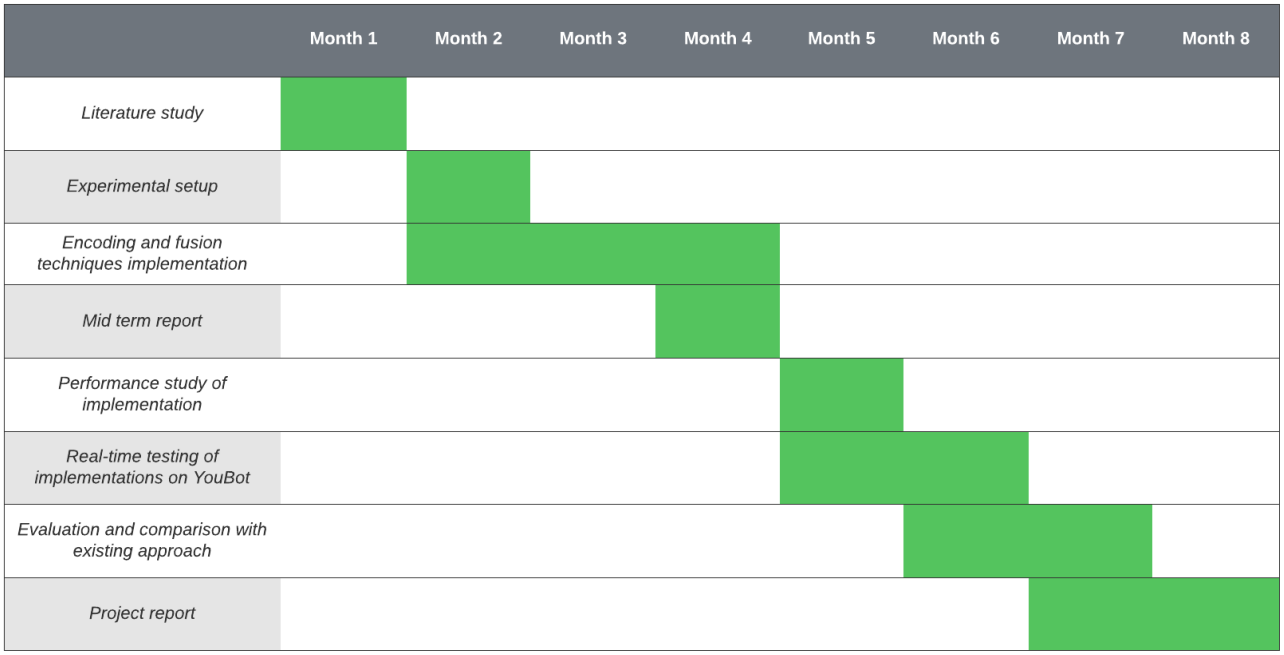

Figure 3: Project schedule

4.4 Deliverables

Minimum Viable

- Literature review insights.
- Training a CGAN model on the RoboCup@Work dataset.
- Implement encoding and fusion techniques.
- R&D final report documentation.

Expected

- Implementation of all the mentioned encoding and fusion techniques.
- Performance and computational analysis of the implemented approach.

Desired

- Real-time testing and integration of implemented approach on the YouBot.
- Comparative analysis of existing YOLOv8 approach vs Implemented approaches.
- Paper publication on the study conducted.

References

[1] George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning, 2020.

[2] Dan Becker. Using categorical data with one hot encoding, 2018. URL `https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding`.

[3] Adam Berger. Error-correcting output coding for text classification. *In Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, 07 2001.

[4] Mwamba Kasongo Dahouda and Inwhee Joe. A deep-learned embedding technique for categorical features encoding. *IEEE Access*, 9:114381–114391, 2021. doi: 10.1109/ACCESS.2021.3104357.

[5] Okwudili M Ezeme, Qusay H. Mahmoud, and Akramul Azim. Design and development of ad-cgan: Conditional generative adversarial networks for anomaly detection. *IEEE Access*, 8:177667–177681, 2020. doi: 10.1109/ACCESS.2020.3025530.

[6] RoboCup Federation. Robocup@work. URL `https://atwork.robocup.org/`.

[7] Ignazio Gallo, Alessandro Calefati, Shah Nawaz, and Muhammad Kamran Janjua. Image and encoded text fusion for multi-modal classification, 2018.

[8] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester*, 2014(5):2, 2014.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[10] Claas Grohnfeldt, Michael Schmitt, and Xiaoxiang Zhu. A conditional generative adversarial network to fuse sar and multispectral optical data for cloud removal from sentinel-2 images. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1726–1729, 2018. doi: 10.1109/IGARSS.2018.8519215.

[11] Jonathan Hui. Gan — cgan infogan (using labels to improve gan), 2018. URL `https://jonathan-hui.medium.com/gan-cgan-infogan-using-labels-to-improve-gan-8ba4de5f9c3d`.

[12] MAS-group. b-it-bots@work, . URL `https://www.h-brs.de/en/a2s/b-it-bots`.

[13] MAS-group. Robocup@work dataset, . URL `https://github.com/b-it-bots/mas_industrial_robotics`.

[14] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022.

[15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[16] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

[17] Priyadarshini Panda, Abhronil Sengupta, and Kaushik Roy. Conditional deep learning for energy-efficient and enhanced pattern recognition, 2016.

[18] Hyojin Park, YoungJoon Yoo, and Nojun Kwak. Mc-gan: Multi-conditional generative adversarial network for image synthesis, 2018.

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

[20] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8, 2023.

[21] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[22] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, August 2023.

[23] Lu Yang. Conditional generative adversarial networks (cgan) for abnormal vibration of aero engine analysis. In *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT*, pages 724–728, 2020. doi: 10.1109/ICCASIT50869.2020.9368622.