Underwater Object Recognition

Tommaso Cammelli

Student ID: 23215488

Faculty of Design and Creative Technologies

Auckland University of Technology

COMP 838: Deep Learning

Dr. Wei Qi Yan

12 April, 2024

Abstract

This project proposal aims to investigates the efficacy of various deep learning architectures in the specialized field of underwater object recognition, a critical area of research with significant implications for marine biology, archaeological exploration, and autonomous underwater navigation. Recognizing the unique challenges posed by underwater environments such as low lighting, unclear images, and complex backgrounds. This project aims to perform a comparative analysis of various deep learning models and architecture like Convolutional Neural Networks, Residual Networks and Transformer models to find the most effective strategies for accurate and efficient object recognition in the underwater realm.

By using Python as the primary programming language, the project will use popular deep learning libraries like TensorFlow and PyTorch for model implementation and training. The methodology focuses on a comprehensive evaluation of model performance, assessing accuracy, precision, recall, and computational efficiency under varying underwater conditions.

Expected outcomes include identifying optimal deep learning architectures that overcome the inherent challenges of underwater imaging.

Underwater Object Recognition

Underwater object recognition has emerged as a crucial field within marine research, underpinning advancements in ecological monitoring, archaeological exploration, and autonomous underwater vehicle navigation. The unique challenges of underwater environments, such as limited visibility, varying light conditions, and the presence of noise, demand robust and adaptive solutions. Deep learning offers a promising solution for addressing these challenges.

Recent years have witnessed the development of various deep learning architectures, each with its strengths and limitations when applied to the underwater domain. Convolutional Neural Networks are well known for using spatial hierarchies for effective feature extraction. However, the introduction of architectures like Residual Networks (ResNets) and Transformer models has opened new possibilities for enhancing recognition accuracy and computational efficiency. This project aims to conduct a comparative analysis of these architectures, evaluating their performance in underwater object recognition tasks.

By analyzing the adaptability of these models to underwater conditions, their scalability, and their efficiency in recognizing a diverse range of objects, this project intends to identify optimal strategies for deep learning-based underwater object recognition. The comparison will not only focus on quantitative performance metrics but also consider factors such as model complexity, training data requirements, and real-world applicability.

The objects to be detected in this study are underwater animal species, including echinus, holothurian, scallop, and starfish. These species have been selected because datasets containing images of these animals are more prevalent and easy to find.

Sections of this proposal

The proposal is divided in the following sections:

- **Literature Review** Detailed examination of previous studies and advancements in underwater object recognition.
- **Methodology** Comprehensive outline of the data collection, model selection,

implementation, and evaluation criteria.

- **Concluding Remarks** Summarization of findings and implications for future research in underwater object recognition.

Literature Review

Underwater object recognition has been subject of increasing interest in the scientific community, driven by the increasing number of potential application in areas such as marine biology, underwater archeology and autonomous underwater navigation.

Deep Learning in Underwater Object Recognition

Deep learning has significantly advanced the field of computer vision, surpassing the capabilities of traditional object detection methods that relied on hand-crafted features for image classification. Initial endeavors predominantly employed Convolutional Neural Networks (CNNs), used for their proficiency in automatically learning and adapting spatial hierarchies of features from images (Zhiqiang & Jun, 2017). For underwater detection CNN were used successfully to recognize objects with significant accuracy, managing to work against the challenges of low-quality images (F. Han et al., 2020)

Advancements and Architectural Innovations

Following the success of CNNs, more sophisticated architectures like Residual Networks (ResNets) have been introduced, allowing training on deeper networks by addressing the vanishing gradient problem (He et al., 2016). ResNets have demonstrated remarkable performance in general object recognition tasks, yet their efficiency and adaptability in underwater conditions remain an area of active research.

The advent of Transformer models, originally designed for natural language processing tasks, has been adapted for computer vision (K. Han et al., 2023). These models, which rely on self-attention mechanisms, offer a new approach to handling the spatial relationships in images, potentially offering advantages in complex underwater scenes where context and object relationships are important.

Challenges in Underwater environments

Recognizing objects underwater presents unique challenges not typically encountered in terrestrial environments. Factors such as variable lighting conditions, water turbidity, and the presence of particulates can significantly impact the performance of deep learning models (Li et al., 2020). Studies have begun to explore the robustness of different architectures under such conditions, emphasizing the need for models that can adapt to or correct for these environmental distortions.

Comparative Analyses in the Literature

Comparative studies specifically addressing the performance of deep learning architectures in underwater object recognition are sparse but growing. These studies are crucial for understanding the practical limitations and opportunities of applying these advanced computational models to underwater scenarios. For instance, a study by Teng and Zhao (2020) compared the accuracy and computational efficiency of CNNs and ResNets in identifying underwater mines, highlighting the trade-offs between model complexity and recognition performance.

Additionally, Pedersen et al. (2019) introduced the Brackish Dataset, a new publicly available underwater dataset containing annotated image sequences of fish, crabs, and starfish captured in brackish water with varying visibility. They evaluated the performance of YOLOv2 and YOLOv3 on this dataset, establishing a baseline for future studies. This contribution is significant as it provides a unique dataset and highlights the importance of robust data for training and evaluating marine object recognition models.

Conclusion

The reviewed literature shows the rapid advancements and diverse applications of deep learning in underwater object recognition. While Convolutional Neural Networks have laid a robust foundation, the emergence of architectures like Residual Networks and Transformer models offers promising avenues for enhancing recognition accuracy and efficiency. However, the unique challenges posed by underwater environments necessitate further research and the development of adaptive models. The introduction

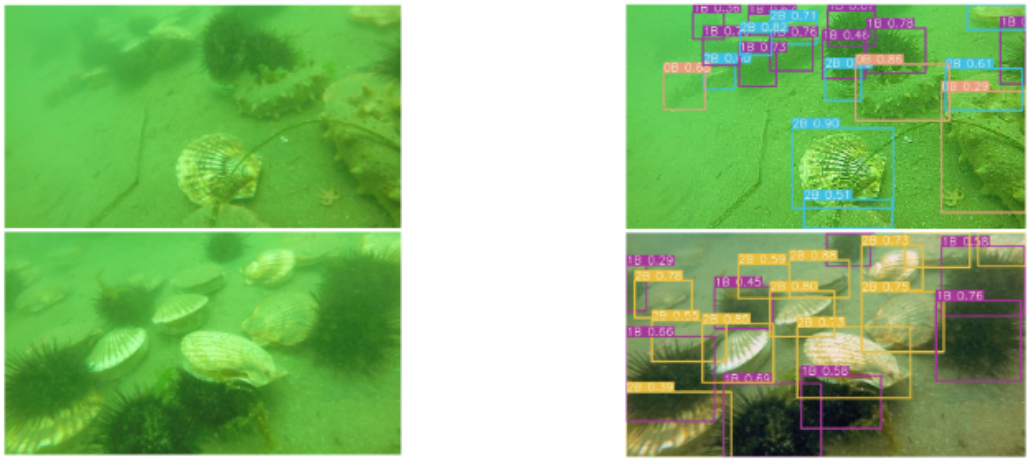

Figure 1

Pre-processing and Post-processing of underwater images (Jiang et al., 2021)

of comprehensive datasets, such as the Brackish Dataset by Pedersen et al. (2019), plays an important role, providing the necessary data for training and evaluation.

Methodology

Data Collection

In this project, publicly available underwater image datasets with labels about underwater species will be used, as these are more prevalent and accessible. Specifically, the datasets to be used include the Underwater Object Detection Dataset (UODD) (Jiang et al., 2021)[1], and the Brackish Dataset introduced by Pedersen et al. (2019) [2]. These datasets contain a diverse range of underwater scenes, including various species such as fish, crabs, and starfish, providing a comprehensive basis for testing and comparison. The annotations are labeled in formats like MS COCO, facilitating consistent and detailed evaluation of model performance.

Model Selection and Development

- **CNNs**: Initial experiments will focus on conventional Convolutional Neural Networks (CNNs), which have proven effective in basic underwater object detection tasks (F. Han et al., 2020). CNNs utilize layers of convolutional filters to extract

[1] UODD dataset is available at https://github.com/LehiChiang/Underwater-object-detection-dataset

[2] Brackish Dataset is available at https://www.kaggle.com/aalborguniversity/brackish-dataset

features from images. Mathematically, a convolution operation on an image I with a filter K is defined as:

$$(I * K)(x, y) = \sum_{i=-m}^{m} \sum_{j=-n}^{n} I(x + i, y + j)K(i, j) \tag{1}$$

where (x, y) are the coordinates of the pixel. These models will serve as a benchmark for comparing more advanced architectures.

- **Residual Networks (ResNets)**: Given their ability to train deeper networks by mitigating the vanishing gradient problem (He et al., 2016), ResNets will be explored for their potential to improve recognition accuracy in complex underwater scenes. ResNets introduce shortcut connections that skip one or more layers, which helps in addressing the vanishing gradient problem. The residual block can be expressed as:

$$y = F(x, \{W_i\}) + x \tag{2}$$

where x is the input to the residual block, F is the residual mapping, and $\{W_i\}$ are the weights of the layers in the block. This allows the network to learn the residual mapping instead of the original unreferenced mapping.

- **Transformers**: The study will also incorporate Transformer models (K. Han et al., 2023), which utilize self-attention mechanisms to examine their effectiveness in capturing the spatial relationships of objects in underwater images. The self-attention mechanism computes a weighted sum of input features, where the weights are determined by the similarity between elements. Given an input sequence of vectors $X = [x_1, x_2, \ldots, x_n]$, the self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q (queries), K (keys), and V (values) are linear transformations of X, and d_k is the dimension of the key vectors. This mechanism allows the model to weigh the importance of different elements in the sequence dynamically.

Implementation

The project will employ Python as the primary programming language, leveraging its extensive ecosystem and the ease of finding pre-implemented models along

with Jupyter notebook to help the visualization of data. For model implementation and training, I will utilize deep learning libraries such as TensorFlow and PyTorch. TensorFlow and PyTorch offer extensive support for building and training deep learning models, including pre-built layers, optimization algorithms, and GPU acceleration.

Experiments will be conducted on a laptop equipped with an Nvidia GPU to facilitate efficient model training and evaluation. The GPU acceleration is critical for handling the computational load of training deep learning models, as it significantly speeds up the process by performing parallel computations.

Evaluation Criteria

The evaluation of the various models will use common indicators like accuracy, precision, recall, and F1 score to determine the effectiveness of each model in correctly identifying and classifying underwater objects.

- **Accuracy**: This measures the overall correctness of the model, defined as the ratio of correctly predicted instances to the total instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- **Precision**: This measures the accuracy of the positive predictions, defined as the ratio of true positive instances to the total predicted positive instances.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

- **Recall**: This measures the ability of the model to identify all relevant instances, defined as the ratio of true positive instances to the total actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

- **F1 Score**: This is the harmonic mean of precision and recall, providing a single metric that balances both. It is defined as:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Additionally, computational efficiency, measured in terms of training time and inference speed, will be considered to assess the practicality of deploying these models in real-world applications. Computational efficiency is crucial as it impacts the feasibility of using these models in environments with limited computational resources or in scenarios requiring real-time processing.

The performance of each architecture will be compared to establish their relative strengths and weaknesses in underwater object recognition. This analysis will also explore the impact of varying dataset complexities and environmental conditions on model performance.

Concluding Remarks

This project proposal outlines a comprehensive approach to exploring the potential of various deep learning architectures in the challenging domain of underwater object recognition. By using Python with the capabilities of TensorFlow, PyTorch, and the interactive environment of Jupyter Notebooks, this study aims to not only compare the efficacy of established models such as CNNs and ResNets but also investigate the emerging potential of Transformer models in this context.

This project could potentially contribute to the field of marine biology, archaeological exploration and autonomous underwater navigation by identifying optimal deep learning strategies that can help with the unique challenges posed by the underwater environments.

References

Han, F., Yao, J., Zhu, H., & Wang, C. (2020). Underwater Image Processing and Object Detection Based on Deep CNN Method. *Journal of Sensors*, *2020*, e6707328. https://doi.org/10.1155/2020/6707328

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Jiang, L., Wang, Y., Jia, Q., Xu, S., Liu, y., Fan, X., Li, H., Liu, R., Xue, X., & Wang, R. (2021). Underwater Species Detection using Channel Sharpening Attention, 4259–4267. https://doi.org/10.1145/3474085.3475563

Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., & Tao, D. (2020). An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Transactions on Image Processing*, *29*, 4376–4389. https://doi.org/10.1109/TIP.2019.2955241

Pedersen, M., Bruslund Haurum, J., Gade, R., & Moeslund, T. B. (2019). Detection of Marine Animals in a New Underwater Dataset with Varying Visibility, 18–26. Retrieved June 9, 2024, from https://openaccess.thecvf.com/content_CVPRW_2019/html/AAMVEM/Pedersen_Detection_of_Marine_Animals_in_a_New_Underwater_Dataset_with_CVPRW_2019_paper.html

Teng, B., & Zhao, H. (2020). Underwater target recognition methods based on the framework of deep learning: A survey. *International Journal of Advanced Robotic Systems*, *17*(6), 1729881420976307. https://doi.org/10.1177/1729881420976307

Zhiqiang, W., & Jun, L. (2017). A review of object detection based on convolutional neural network. *2017 36th Chinese Control Conference (CCC)*, 11104–11109. https://doi.org/10.23919/ChiCC.2017.8029130

Appendix

Responses to Reviewers Comments on the Project Proposal

Reviewer 1

Question 1.1: Consider may adding more detail to the introduction on the types of objects you will be detecting underwater. It is mentioned in the methodology but I believe that it would be good to know that in the introduction as well.

Answer 1.1: Updated the introduction adding type of objects that will be detected (Animal Species).

Question 1.2: It would be helpful if the literature review includes a conclusion. Having a conclusion will help get summarised results and some key findings of the literature review.

Answer 1.2: Updated the literature review section adding a conclusion.

Reviewer 2

Question 2.1: You could add some details about the dataset you are using, and some illustrations so we could understand easier why you are using this dataset.

Answer 2.1: Added details about the used dataset in both introduction and methodology.

Question 2.2: You should be more understandable in your methodology part. You could add some mathematical explanations about the concepts you talked about

Answer 2.2: Added more details about the models and concepts used in the methodology section along with formulas.

Reviewer 3

Question 3.1: Explain the structure of the dataset and the model. Is it zoomed in on a particular element and then labelled? For the dataset, you should probably explain in detail the different features you are using. For the model, just make a scheme of your project's architecture.

Answer 3.1: Updated details about the dataset and the model in the methodology section.

Question 3.2: Instead of focusing on the confusion matrix equations themselves, talk about their practical application within your proposal. I would use the equations to detail some parts of the model you are working on. The transformative models, for example seem really interesting to study mathematically.

Answer 3.2: Updated validation criteria and added more details about each model.

Question 3.3: Downsizing the repetition of the concepts is not an easy task. However, your project might benefit from finding alternative ways to convey this information. In the case mentioned above, I would maybe detail one specific use-case like archaeology and explain how it can benefit from what you are studying.

Answer 3.3: Restructured text to limit repetition in all sections.