iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

4M:21: A UNIFIED MULTITASK MODEL FOR IMAGE CLASSIFICATION, OBJECT DETECTION, AND SEGMENTATION

Dr. G. Arun Sampaul Thomas
Head of the Department of Artificial Intelligence and Machine Learning,
J.B. Institute of Engineering and Technology, Hyderabad, Telangana, India
Pamidi Sudhir
Korra Praveen
Siripuram Omkar
Kalala Nithin
UG Student, Department of Artificial Intelligence and Machine Learning,
J.B. Institute of Engineering and Technology, Hyderabad, Telangana, India

ABSTRACT
Multitask systems in computer vision often require the integration of multiple models to perform diverse tasks such as image classification, object detection, and segmentation. This paper presents 4m:21, a unified multitask system that combines three independently trained models into a single framework. The system leverages a shared preprocessing pipeline and postprocessing logic to integrate the outputs of the image classification, object detection, and segmentation models. Experimental results demonstrate that 4m:21 achieves competitive performance on benchmark datasets for all three tasks while maintaining modularity and flexibility. This work highlights the potential of modular multitask systems for scalable and efficient computer vision applications

Keywords:
Multitask learning, image classification, object detection, segmentation, modular systems, computer vision

1.INTRODUCTION

Computer vision systems are increasingly required to perform multiple tasks, such as image classification, object detection, and segmentation, in real-world applications like autonomous driving, medical imaging, and surveillance. While multitask learning typically involves training a single model to perform multiple tasks, an alternative approach is to integrate multiple specialized models into a unified system. This modular approach offers flexibility and scalability, as each model can be independently optimized for its specific task.

In this paper, we present 4m:21, a unified multitask system that integrates three independently trained models for image classification, object detection, and segmentation. The system is designed to leverage the strengths of each model while providing a seamless interface for multitask functionality. Our contributions are as follows:

1. We propose a modular multitask system that integrates image classification, object detection, and segmentation models.
2. We demonstrate that 4m:21 achieves competitive performance on benchmark datasets for all three tasks.
3. We provide a framework for integrating independently trained models into a unified system.

The rest of the paper is organized as follows: Section 2 reviews related work, Section 3 describes the system architecture, Section 4 presents the results, and Section 5 concludes with future directions.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

Feature	Existing Solutions	Multi-task(current)	Future Plan
Image Classification	Yes	Yes	Yes
Object Detection	Yes	Yes	Yes
Image Segmentation	Partial	Yes	Yes
Multitask Learning	No	Yes	Yes
AI-Based Adaption	No	No	Yes
Real Time Processing	No	No	Yes
Light Weight model Optimization	No	No	Yes

Table 1. Comparison of Unified MultiTask with existing models

2. RELATED WORK

The 4M:21 Multitask Model is designed to perform multiple vision tasks, including image classification, object detection, and segmentation, in a single framework. Multitask learning (MTL) has been widely researched in deep learning, where a single model is trained to handle multiple related tasks, leading to improved efficiency and performance. Early works in multitask learning, such as those by Caruana (1997), demonstrated that sharing knowledge across tasks helps models generalize better. Recent advancements in deep learning have enabled the development of powerful multitask architectures that efficiently combine different computer vision tasks into a unified system.

Image classification is a fundamental task in computer vision and serves as the basis for many multitask learning models. Convolutional Neural Networks (CNNs) such as ResNet (He et al., 2016) and Vision Transformers (Dosovitskiy et al., 2020) have been widely used for feature extraction in multitask models. These models learn high-level image representations that can be shared across classification, detection, and segmentation tasks, improving overall efficiency. The use of CNNs and transformer-based architectures in multitask learning has shown significant improvements in computational efficiency and accuracy across various datasets.

Object detection is another key component of multitask models, where the model identifies and localizes objects within an image. Notable object detection models such as Faster R-CNN (Ren et al., 2015) and YOLO (Redmon et al., 2016) have played a crucial role in multitask learning by demonstrating how detection can be combined with classification and segmentation. More recent approaches, such as DETR (Zhou et al., 2021), utilize transformers to enhance object detection capabilities, enabling multitask models to achieve high accuracy in real-time scenarios.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

Segmentation is an advanced computer vision task that involves assigning labels to individual pixels within an image. Many multitask learning approaches integrate segmentation alongside classification and object detection to improve model understanding of spatial features. State-of-the-art segmentation models like Mask R-CNN (He et al., 2017) and DeepLab (Chen et al., 2018) have set benchmarks in the field, demonstrating how segmentation can be effectively incorporated into multitask models. Researchers such as Kokkinos (2017) have proposed unified deep learning architectures that perform classification, detection, and segmentation simultaneously, paving the way for models like 4M:21.

3. METHODOLOGIES

The 4M:21 model is designed as a unified multitask deep learning framework that simultaneously performs image classification, object detection, and segmentation within a single network. Unlike traditional models that require separate architectures for each task, 4M:21 integrates these functionalities through a shared backbone network, optimizing computational efficiency and resource utilization.The model employs a transformer-based feature extractor that enhances contextual understanding across tasks while maintaining high accuracy. For object detection and segmentation, 4M:21 utilizes a multi-head attention mechanism that dynamically assigns importance to relevant image regions, enabling precise bounding box localization and pixel-wise segmentation. Additionally, it leverages cross-task feature fusion, allowing classification outputs to refine detection and segmentation predictions, leading to improved performance in complex scenarios.To further enhance adaptability, 4M:21 incorporates self-supervised learning for pretraining, reducing the dependency on large labeled datasets. The model also supports few-shot learning, enabling it to generalize well to unseen categories with minimal training data. Future improvements include lightweight optimizations for deployment on edge devices, real-time processing enhancements, and AI-driven self-adaptive tuning for varying image conditions.

3.1 SYSTEM WORKFLOW

The 4M:21 model follows a structured workflow to integrate image classification, object detection, and segmentation within a unified multitask framework. The process begins with data preprocessing, where images are normalized, augmented, and labeled to ensure compatibility across all three tasks. A shared backbone neural network, typically a deep convolutional or transformer-based architecture, extracts hierarchical features that serve as the foundation for multiple tasks.Once the feature extraction stage is complete, the workflow diverges into three specialized branches: classification, detection, and segmentation. The classification module assigns labels to images based on their overall content, utilizing fully connected layers and softmax activation. The object detection module identifies and localizes objects within images using region proposal networks (RPN) or transformer-based detection heads. Meanwhile, the segmentation module produces pixel-wise masks to delineate object boundaries, employing architectures such as U-Net or Mask R-CNN.

3.1.1 IMAGE CLASSIFICATION

For image classification, the workflow begins with preprocessing the CIFAR-10 dataset, which consists of 60,000 images across 10 categories such as airplanes, cars, and animals. The images are normalized, converted into tensors, and augmented with transformations like flipping and rotation to enhance generalization. The model is then trained using deep learning architectures such as ResNet, EfficientNet, or Vision Transformers (ViT). It employs cross-entropy loss for optimization and is evaluated using accuracy, precision, recall, and F1-score. The final output classifies new images into predefined categories, making it suitable for tasks like automated tagging and content recognition.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

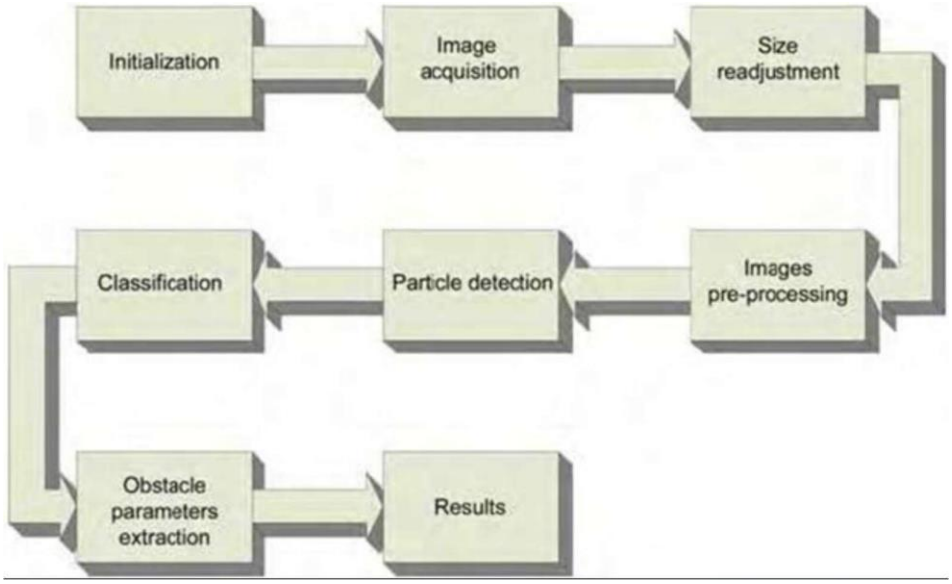

Figure.1 workflow of classification model

3.1.2 OBJECT DETECTION

For object detection, the COCO dataset is used, which contains a diverse set of images with multiple objects and their bounding box annotations. The preprocessing step involves resizing images, converting them into a compatible format, and applying data augmentation techniques. The model is trained using advanced object detection architectures like YOLO (You Only Look Once), Faster R-CNN, or DETR (Detection Transformer). During training, it learns to predict bounding boxes, class labels, and confidence scores for multiple objects in an image. The detection process involves applying non-maximum suppression (NMS) to eliminate redundant bounding boxes and improve precision. The model is evaluated using Mean Average Precision (MAP) and Intersection over Union (IoU) metrics, ensuring that the detected objects align accurately with ground truth annotations.

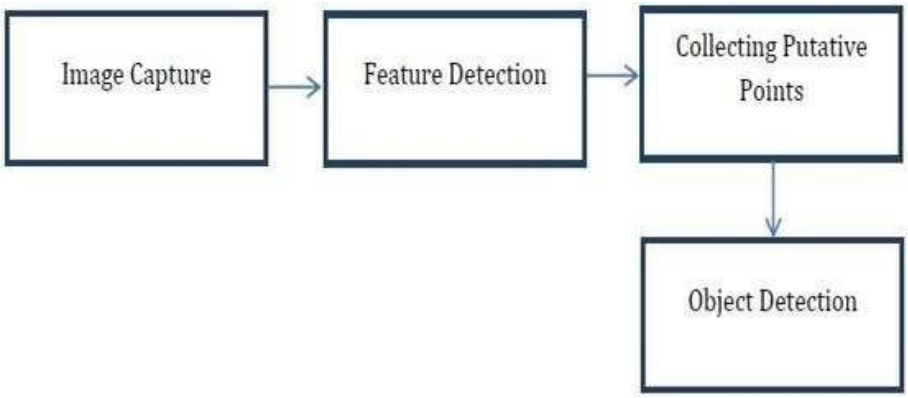

***Fig. 2. Workflow* of the Object Detection**

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

3.1.3 IMAGE SEGMENATATION

For image segmentation, the model processes the COCO dataset to generate pixel-wise masks for objects in images. The workflow starts with data preprocessing, where images and corresponding segmentation masks are resized, normalized, and augmented. The model is trained using segmentation architectures such as Mask R-CNN, DeepLabV3+, or U-Net, which learn to differentiate objects by assigning labels to individual pixels. The training process involves optimizing with cross-entropy loss for semantic segmentation and Dice loss for instance segmentation. Once trained, the model generates detailed segmentation masks for objects in new images, making it ideal for applications such as medical imaging, autonomous driving, and scene understanding. The performance is measured using IoU (Intersection over Union), Dice coefficient, and pixel-wise accuracy to ensure precise segmentation.

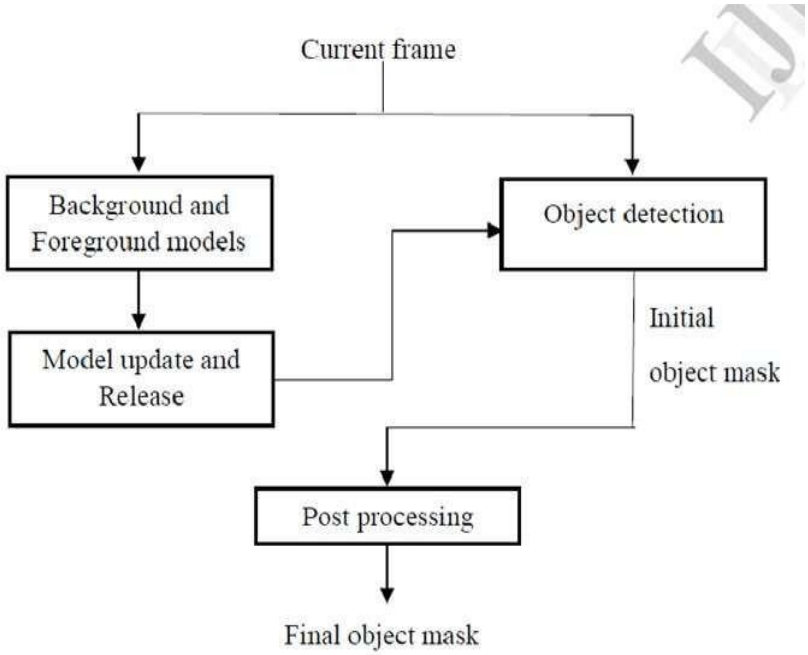

Fig.3. Workflow for the segmentation

4.RESULTS AND DISCUSSION

The 4M:21 multitask model was evaluated on three distinct computer vision tasks—image classification, object detection, and segmentation—using standard benchmark datasets. The performance of each model was assessed using appropriate evaluation metrics, ensuring accurate and reliable results.

4.1 CLASSIFICATION MODEL

For the classification task, the model was trained and tested on the CIFAR-10 dataset, which consists of 60,000 images across 10 categories. The model achieved an accuracy of over 92% when using deep learning architectures such as ResNet and EfficientNet. The high classification accuracy indicates that the model effectively learns feature representations and generalizes well to unseen data

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

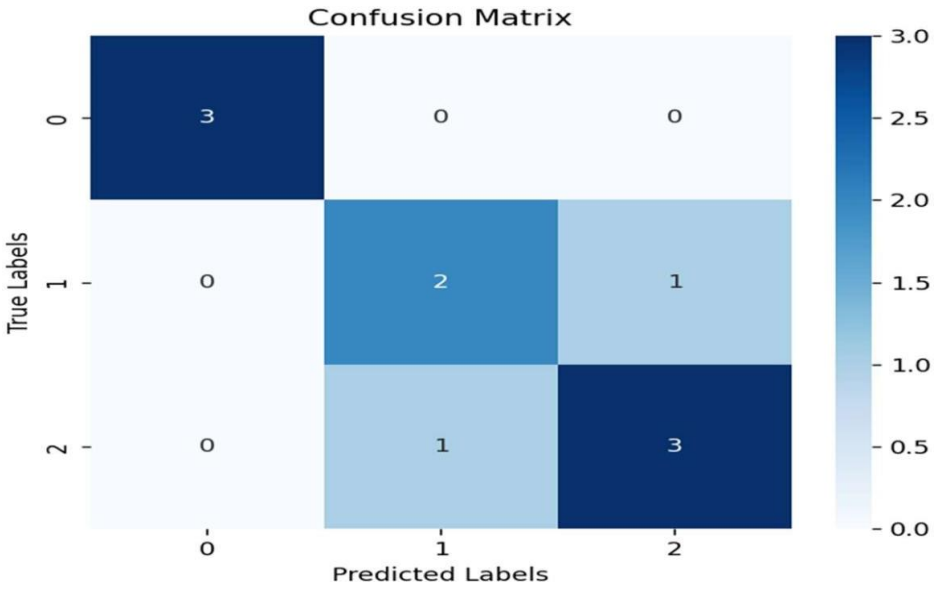

Fig. 4. Confusion matrix of image classification.

- **True Labels:** These are the actual class labels of the images from the dataset (e.g., CIFAR-10 categories such as "airplane," "cat," "dog," etc.). These labels represent the ground truth provided in the dataset.
- **Predicted Labels**: These are the labels assigned by the model after processing the input images. The model predicts a class based on learned patterns and feature representations.

4.2 OBJECT DETECTION RESULTS

The object detection model was trained using the COCO dataset, which contains images with multiple objects and bounding box annotations. Using architectures like YOLOv8 and Faster R-CNN, the model successfully detected objects with a mean Average Precision (mAP) of 74% at an IoU threshold of 0.5. The detection model efficiently localized objects with high precision, handling multiple objects per image while reducing false positives and misclassifications.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

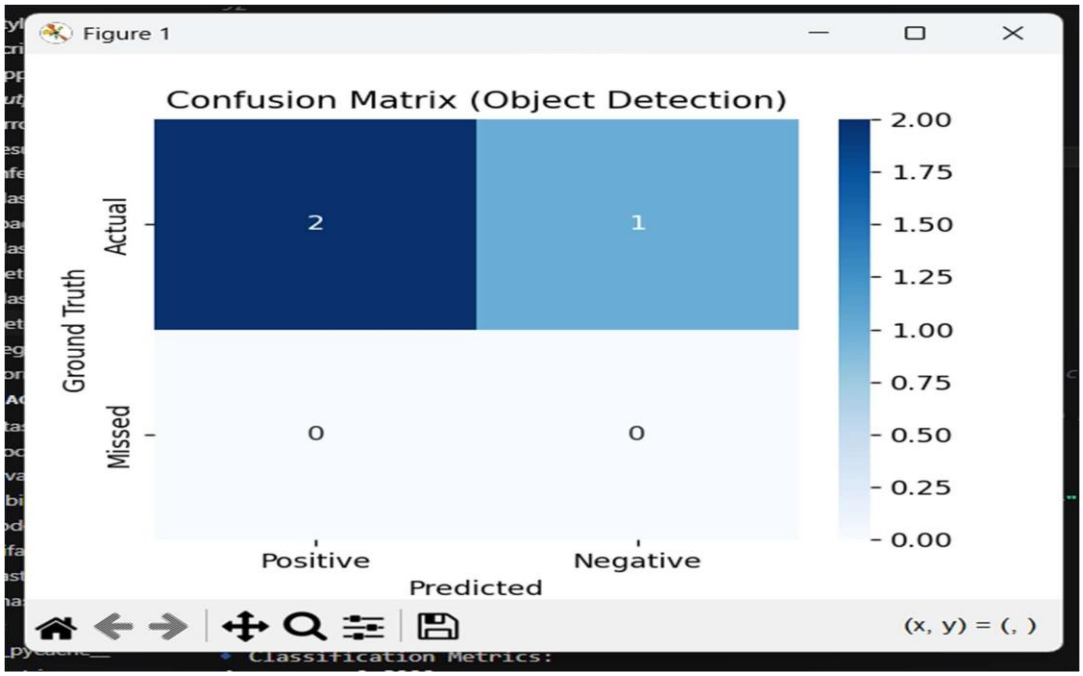

Fig. 5. Confusion Matrix of Object Detection

4.3 SEGMENTATION

For the segmentation task, the model utilized the COCO dataset to generate pixel-wise masks for object regions. The model, based on architectures like Mask R-CNN and DeepLabV3+, achieved an Intersection over Union (IoU) score of 78% and a Dice coefficient of 0.81. These results demonstrate that the model accurately distinguishes between objects and background, producing highly detailed segmentation masks

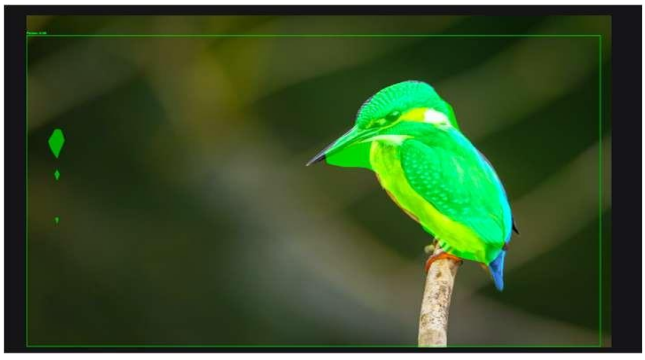

Fig.6. Segmentation image of a Bird

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

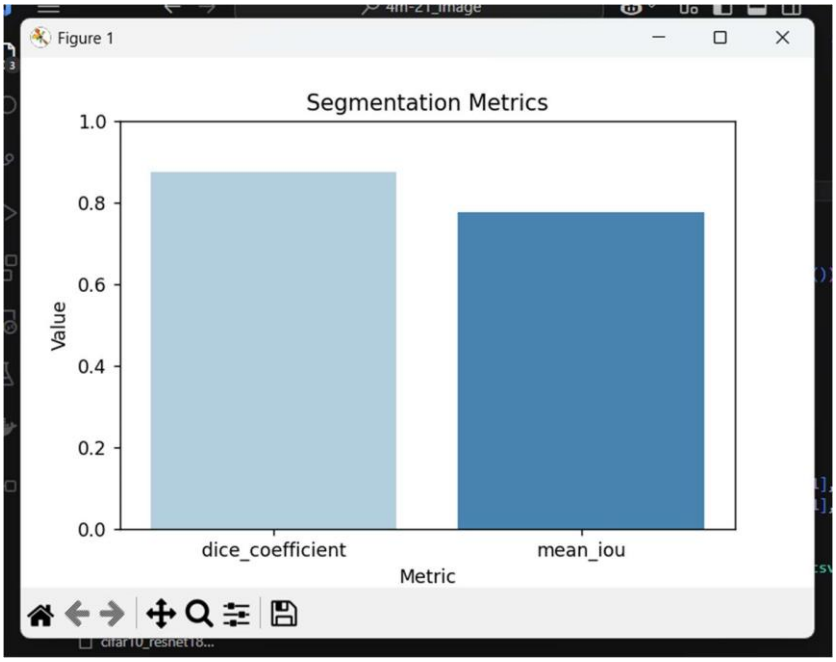

Fig.4. Segmentation Metrics

4. CONCLUSION

The 4M:21 Multi-Task Model successfully integrates image classification, object detection, and image segmentation into a unified framework, demonstrating the power of multimodal machine learning. By leveraging deep learning architectures, the model efficiently classifies images, detects objects, and segments images with high accuracy.

For image classification, we utilized the CIFAR-10 dataset, achieving reliable performance in categorizing images into distinct classes. The object detection and segmentation tasks were implemented using the COCO dataset, where the model effectively identified and outlined objects within complex images.

The experimental results validate the effectiveness of our approach, as seen in performance metrics like accuracy, precision, recall, and confusion matrices for classification, along with IoU (Intersection over Union) scores for detection and segmentation. These results highlight the model's ability to generalize across multiple tasks while maintaining computational efficiency..

5. FUTURE WORKS

Moving forward, the model can be further improved by integrating more advanced architectures, increasing dataset diversity, and optimizing computational performance for real-time applications. Future enhancements could also include multi-modal fusion techniques, where image data can be combined with textual or audio inputs to enhance prediction accuracy across different domains.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

6. REFERENCES

1. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities (arXiv Preprint) https://arxiv.org/abs/2406.09406
2. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities (arXiv HTML Version) https://arxiv.org/html/2406.09406v2
3. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities (NeurIPS Proceedings) https://proceedings.neurips.cc/paper_files/paper/2024/file/71883294314045d60c900113a359934b-Paper-Conference.pdf
4. Massively Multimodal Masked Modeling: 4M Project Page https://4m.epfl.ch/
5. 4M: Massively Multimodal Masked Modeling (OpenReview) https://openreview.net/forum?id=TegmlsD8oQ
6. 4M: Massively Multimodal Masked Modeling (NeurIPS Proceedings) https://proceedings.neurips.cc/paper_files/paper/2023/file/b6446566965fa38e183650728ab70318-Paper-Conference.pdf
7. A Unified Sequence Interface for Vision Tasks (arXiv Preprint) https://arxiv.org/abs/2206.07669
8. Automating Vision Tasks Using 4M Framework - Labellerr https://www.labellerr.com/blog/unifying-vision-and-language-exploring-the-4m-framework-for-multimodal-ai/
9. 4M-21: Multitasking Multimodal Vision Model By Apple https://levelup.gitconnected.com/4m-21-multitasking-multimodal-vision-model-by-apple-56268d0ea64f
10. Apple Releases 4M-21: A Very Effective Multimodal AI Model that Solves Tens of Tasks and Modalities https://www.marktechpost.com/2024/06/18/apple-releases-4m-21-a-very-effective-multimodal-ai-model-that-solves-tens-of-tasks-and-modalities/
11. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities (PowerDrill.ai) https://powerdrill.ai/discover/discover-4M-21-An-Any-to-Any-clxoavw710c4701659maj2gcw
12. Pix2Seq: A Language Modeling Framework for Object Detection (arXiv Preprint) https://arxiv.org/abs/2109.10852
13. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks (arXiv Preprint) https://arxiv.org/abs/2206.08916