MUHAMMAD MAAZ

muhammad.maaz@mbzuai.ac.ae, +971-52-5326156
Website: mmaaz60.github.io, GitHub: mmaaz60 LinkedIn: mmaaz60

PERSONAL PROFILE

Computer Vision engineer with hands-on experience in design, engineering, deployment and monitoring phases of Deep Learning driven Computer Vision products. Aim to work in an organization which provides the opportunity to improve skills, vision and knowledge to grow along with the organization objective.

RESEARCH PROJECTS

Class-agnostic Object Detection with Multi-modal Transformer *Mar 2022 (ECCV-2022)*

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Khan, Rao M. Anwer, Ming-Hsuan Yang

In this work, we explore the potential of the recent Multi-modal Vision Transformers (MViTs) for class-agnostic object detection. Our extensive experiments across various domains and novel objects show the state-of-the-art performance of MViTs to localize generic objects in images. We also develop an efficient and flexible MViT architecture using multi-scale feature processing and deformable self-attention that can adaptively generate proposals given a specific language query.

Bridging the Gap between Object and Image-level Representations for Open-Vocabulary Detection *May 2022 (NeurIPS-2022)*

Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, Fahad Khan

In this work, we propose to solve the Open-vocabulary detection (OVD) problem using pretrained CLIP model, adapting it for object-centric local regions using region-based distillation and image-level weak supervision. Specifically, we propose to utilize high-quality class-agnostic and class-specific object proposals via the pre-trained mulit-modal vision transformers (MViT). The class-agnostic proposals are used to distill region-specific information from CLIP and class-specific proposals allows us to visually ground large vocabularies. We also introduce a region-conditioned weight transfer method to get complementary benefits from both region-based distillation and image-level supervision.

EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications *Jul 2022 (CADL, ECCVW-2022)*

Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, S. Waqas Zamir, Rao M. Anwer, Fahad Khan

In this work, we designed resource-efficient general purpose backbone network for vision tasks. We combine the strengths of both CNN and Transformer models and propose a new efficient hybrid architecture EdgeNeXt. Specifically in EdgeNeXt, we introduce split depth-wise transpose attention (SDTA) encoder that splits input tensors into multiple channel groups and utilizes depth-wise convolution along with self-attention across channel dimensions to implicitly increase the receptive field and encode multi-scale features. Our extensive experiments on classification, detection and segmentation tasks, reveal the merits of the proposed approach, outperforming state-of-the-art methods with comparatively lower compute requirements.

UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation *Dec 2022 (Under review)*

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, Fahad Khan

In this work, we propose a 3D medical image segmentation approach, named UNETR++, that offers both high-quality segmentation masks as well as efficiency in terms of parameters and compute cost. The core of our design is the introduction of a novel efficient paired attention (EPA) block that efficiently learns spatial and channel-wise discriminative features using a pair of inter-dependent branches based on spatial and channel attention. Our spatial attention formulation is efficient having linear complexity with respect to the input sequence length. To enable communication between spatial and channel-focused branches, we share the weights of query and key mapping functions that provide a complimentary benefit (paired attention), while also reducing the overall network parameters.

Fine-tuned CLIP Models are Efficient Video Learners *Dec 2022 (Under review)*

Hanoona Rasheed, Muhammad Uzair Khattak, ***Muhammad Maaz****, Salman Khan, Fahad Khan*

In this work, we show that a simple Video Fine-tuned CLIP (ViFi-CLIP) baseline is generally sufficient to bridge the domain gap from images to videos. Our qualitative analysis illustrates that the frame-level processing from CLIP image-encoder followed by feature pooling and similarity matching with corresponding text embeddings helps in implicitly modeling the temporal cues within ViFi-CLIP. Such fine-tuning helps the model to focus on scene dynamics, moving objects and inter-object relationships. For low-data regimes where full fine-tuning is not viable, we propose a 'bridge and prompt' approach that first uses fine-tuning to bridge the domain gap and then learns prompts on language and vision side to adapt CLIP representations.

MaPLe: Multi-modal Prompt Learning *Dec 2022 (Under review)*

Muhammad Uzair Khattak, Hanoona Rasheed, ***Muhammad Maaz****, Salman Khan, Fahad Khan*

In this work, we propose to learn prompts in both vision and language branches of pretrained CLIP for adapting it to different downstream tasks. Previous works only use prompting in either language or vision branch. We note that using prompting to adapt representations in a single branch of CLIP (language or vision) is sub-optimal since it does not allow the flexibility to dynamically adjust both representation spaces on a downstream task. To this end, we propose Multi-modal Prompt Learning (MaPLe) for both vision and language branches to improve alignment between the vision and language representations. Our design promotes strong coupling between the vision-language prompts to ensure mutual synergy and discourages learning independent uni-modal solutions.

EDUCATION

Mohamed bin Zayed University of Artificial Intelligence, UAE *Dec 2020 - Dec 2022*

Research Based Masters in Computer Vision

CGPA: 4.0/4.0

University of Engineering and Technology, Pakistan *Sep 2014 - Aug 2018*

B.Sc. Electrical Engineering

CGPA: 3.7/4.0 (First class with honors)

WORK EXPERIENCE

Hazen.ai *Jul 2020 - Dec 2020*

Computer Vision Engineer

Developed a traffic light phase detection solution for road safety applications. I trained a network to learn embeddings for traffic light phases (red, yellow, green and black) using triplet loss. The network was robust enough to handle different road scenarios including day and night scenes. The product was deployed on the NVIDIA Jetson devices using TensorRT.

Confiz Limited *Jun 2018 - Jul 2020*

Computer Vision Engineer

Led Shopper Value - Computer Vision Team where I was responsible for technological evolution and scalability of Computer Vision Products; Visitor Tracking and Visitor Profile.

- **Visitor Tracking:** A Person Detection and Tracking solution to identify the engaged and ignored areas of a retail store. Our utmost challenge was to process 7 to 10 video streams on an i5 CPU or NVIDIA Jetson device with fair enough accuracy. We experimented with Yolov3 and pruned it to get the desired speed and accuracy balance. We used Network Distillation to train camera specific small neural networks. We also focused to effectively utilize the CPU cores and use optimized inference frameworks like Intel's OpenVino and TensorRT for edge deployment.
- **Visitor Profile:** A face recognition solution capable of generating visitor's and buyer's demographics and visit frequency data for the brick and mortar retail stores. FaceNet like architecture was being used to prepare face embeddings.

Mentor, A Siemens Buisness *Jun 2017 - Aug 2017*
Software Engineer

During the stay, I developed small multithreaded applications, created Linux distribution following the LFS document, built customized Embedded Linux for Raspberry Pi using Yocto project, learned about cross-compilation and wrote a GPIO device driver for Raspberry Pi embedded board.

CERTIFICATES

Computer Vision Nano Degree *Udacity*

Deep Learning Specialization by deeplearning.ai *Coursera*

Machine Learning with TensorFlow on Google Cloud Platform Specialization *Coursera*

Advance Machine Learning with TensorFlow on Google Cloud Platform Specialization *Coursera*

TECHNICAL STRENGTHS

Computer Sciences	Computer Vision, Deep Learning, Machine Learning
Programming Languages	Python, C, Java
Softwares & Tools	Pycharm, VS Code, MATLAB
ML and DL Frameworks	PyTorch, TensorFlow (basics only)

EXTRA-CURRICULAR

- Former Secretary of Graduate Student Council at MBZUAI
- Former Assistant Vice President Operation at IET UET Chapter
- Member of Education for Every Child (EFE) foundation
- Enjoy travelling, cricket and table tennis

REFERENCES

Dr. Salman Khan
Academic Advisor
Associate Professor at MBZUAI
✉ salman.khan@mbzuai.ac.ae

Dr. Fahad Khan
Academic Advisor
Associate Professor at MBZUAI
✉ fahad.khan@mbzuai.ac.ae

Mr. Hashim Ali
Chief Operating Officer
Confiz Limited
✉ hashim.ali@confiz.com