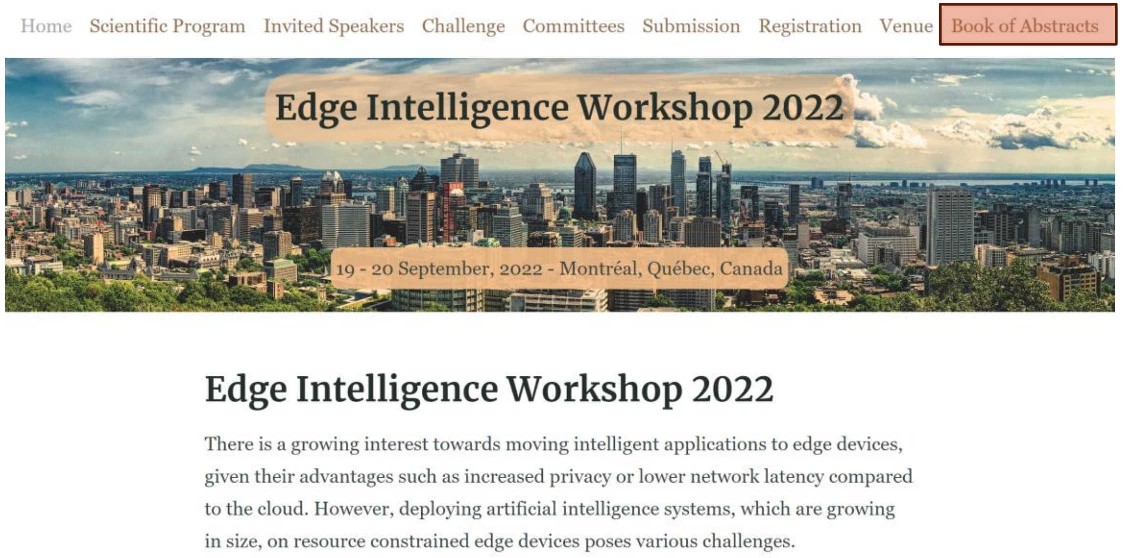
# EDGE INTELLIGENCE OPENING

AAAI-EIW-III

https://eiw2024.github.io/

# HISTORY



Home Scientific Program Invited Speakers Challenge Committees Submission Registration Venue Book of Abstracts

Edge Intelligence Workshop 2022

19 - 20 September, 2022 - Montréal, Québec, Canada

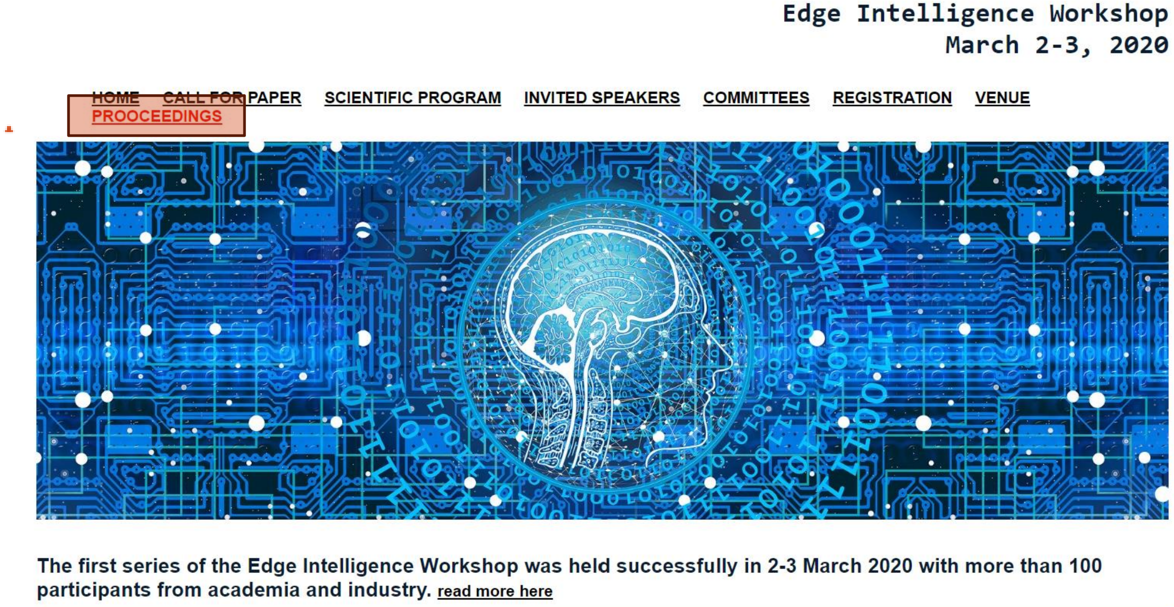
## Edge Intelligence Workshop 2022

There is a growing interest towards moving intelligent applications to edge devices, given their advantages such as increased privacy or lower network latency compared to the cloud. However, deploying artificial intelligence systems, which are growing in size, on resource constrained edge devices poses various challenges.

https://eiw2022.github.io/

https://eiw2022.github.io/assets/Proceedings.pdf

https://eiw2024.github.io/



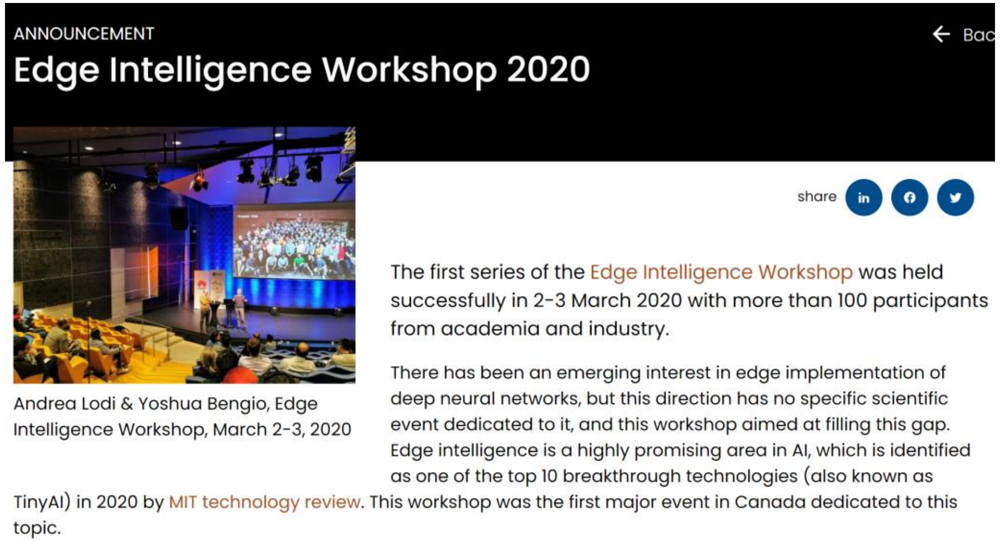
Edge Intelligence Workshop
March 2-3, 2020

HOME CALL FOR PAPER SCIENTIFIC PROGRAM INVITED SPEAKERS COMMITTEES REGISTRATION VENUE
PROOCEEDINGS

The first series of the Edge Intelligence Workshop was held successfully in 2-3 March 2020 with more than 100 participants from academia and industry. read more here

https://www.gerad.ca/colloques/EdgeIntelligence2020/

https://www.gerad.ca/en/papers/G-2020-23



ANNOUNCEMENT

← Bac

## Edge Intelligence Workshop 2020

share

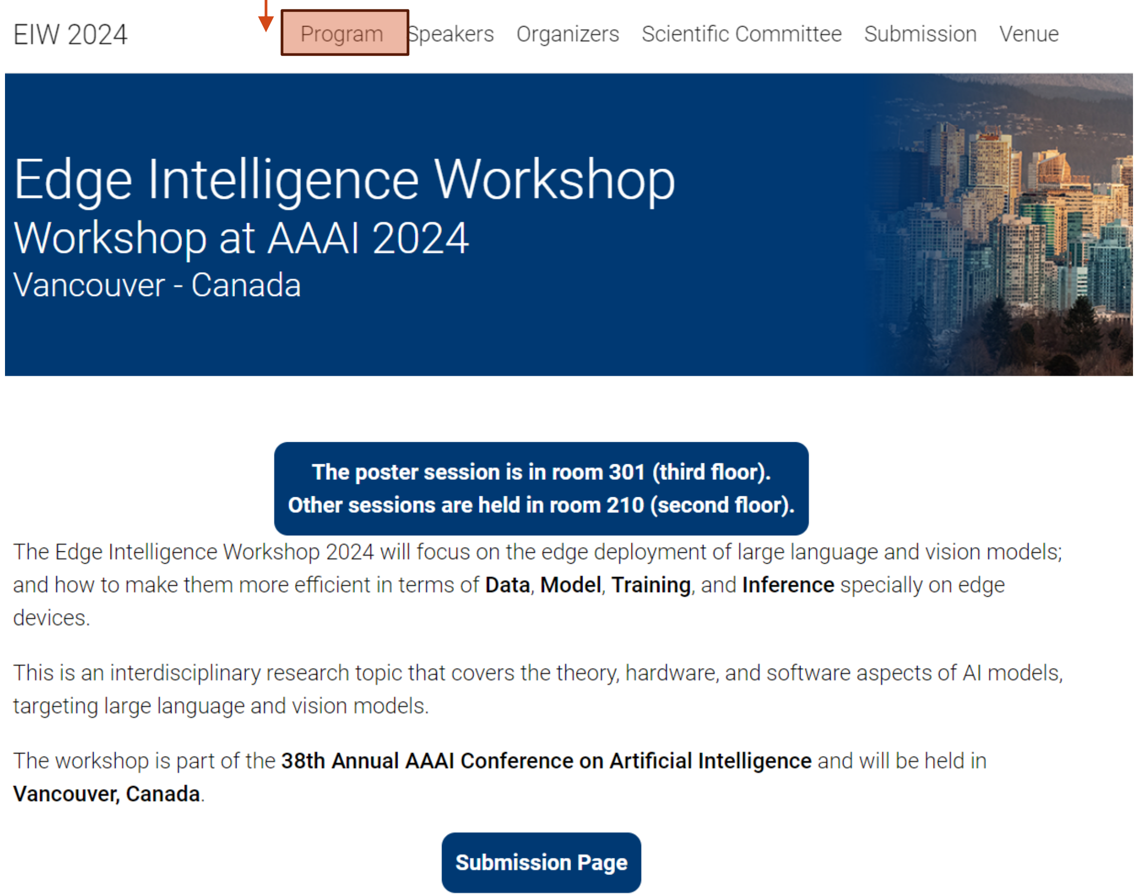
Andrea Lodi & Yoshua Bengio, Edge Intelligence Workshop, March 2-3, 2020

The first series of the Edge Intelligence Workshop was held successfully in 2-3 March 2020 with more than 100 participants from academia and industry.

There has been an emerging interest in edge implementation of deep neural networks, but this direction has no specific scientific event dedicated to it, and this workshop aimed at filling this gap. Edge intelligence is a highly promising area in AI, which is identified as one of the top 10 breakthrough technologies (also known as TinyAI) in 2020 by MIT technology review. This workshop was the first major event in Canada dedicated to this topic.

# WEBSITE AND SLIDES

https://eiw2024.github.io/



Edge Intelligence Workshop
Workshop at AAAI 2024
Vancouver - Canada

**The poster session is in room 301 (third floor).**
**Other sessions are held in room 210 (second floor).**

The Edge Intelligence Workshop 2024 will focus on the edge deployment of large language and vision models; and how to make them more efficient in terms of **Data**, **Model**, **Training**, and **Inference** specially on edge devices.

This is an interdisciplinary research topic that covers the theory, hardware, and software aspects of AI models, targeting large language and vision models.

The workshop is part of the **38th Annual AAAI Conference on Artificial Intelligence** and will be held in **Vancouver, Canada**.

https://eiw2024.github.io/

# ORGANIZERS

Warren Gross

McGill University

Vahid Partovi Nia

Polytechnique Montreal, Huawei Noah's Ark Lab

Andrea Lodi

Cornell Tech

Shahrokh Valaee

University of Toronto

Melika Payvand

UZH and ETH Zurich

Mehdi Rezagholizadeh

Huawei Noah's Ark Lab

Habib Hajimolahoseini

Huawei Toronto Research Centre

Mouloud Belbahri

Layer6AI TDBank

Mohammadreza Tayaranian

McGill University

Yuanhao Yu

Huawei Noah's Ark Lab

Ali Edalati

Huawei Noah's Ark Lab

# SCIENTIFIC COMMITTEE

| | |
|---|---|
| Abbas Ghaddar | Huawei Noah's Ark Lab |
| Alireza Ghaffari | Huawei Noah's Ark Lab |
| Ali Edalati | Huawei Noah's Ark Lab |
| Didier Chételat | Polytechnique Montreal |
| Ghouthi Boukli-Hacene | Sony |
| Gonçalo Mordido | MILA |
| Erfan Seyedsalehi | Huawei Noah's Ark Lab |
| Ehsan Kamalloo | University of Alberta |
| Hassan Mozafari | McGill University |
| Khalil Bibi | Huawei Noah's Ark Lab |
| Marzieh Tahaei | Huawei Noah's Ark Lab |
| Michael Metel | Huawei Noah's Ark Lab |
| Mohammadreza Tayaranian | McGill University |
| Mojtaba Valipour | University of Waterloo |
| Peng Lu | Université de Montreal |
| Sharareh Younesian | Huawei Noah's Ark Lab |
| Ramchalam Ramakrishnan | Qualcomm |
| Ritam Haldar | Qualcomm |
| Vanessa Courville | Huawei Noah's Ark Lab |
| Walid Ahmed | Huawei Toronto Research Centre |
| Xinlin Li | Huawei Noah's Ark Lab |
| Zhixiang Chi | Huawei Noah's Ark Lab |

# KEYNOTE SPEAKERS

Diana Marculescu

University of Texas at Austin

Sarath Chandar

Polytechnique Montreal, MILA

Lukas Cavigelli

Huawei Zurich Research Center

Pascal Poupart

University of Waterloo

Di Niu

University of Alberta

# SCHEDULE

| Start Time | End Time | Program | Speaker |
|---|---|---|---|
| 08:45 | 09:00 | Opening Remarks | |
| 09:00 | 09:30 | Invited Talk | Diana Marculescu |
| 09:30 | 10:00 | Nominated Papers | Diana Marculescu,<br>Mehdi Rezagholizadeh |
| 10:00 | 10:30 | Invited Talk | Di Niu |
| 10:30 | 11:00 | Coffee Break | |
| 11:00 | 11:30 | Invited Talk | Pascal Poupart |
| 11:30 | 12:30 | Poster Discussion & Break | |
| 12:30 | 14:00 | Lunch | |
| 14:00 | 15:00 | Panel Discussion | |
| 15:00 | 15:30 | Invited Talk | Sarath Chandar |
| 15:30 | 16:00 | Coffee Break | |
| 16:00 | 16:30 | Invited Talk | Lukas Cavigelli |
| 16:30 | 16:45 | Closing Remarks | |

# POSTERS 11:30-12:30 THEN LUNCH

- Room 301
- Panel numbers 62-76
- Lunch 12:30-14:00

# PANEL DISCUSSION AFTER LUNCH 14:00-15:00

Mehdi Rezagholizadeh

Huawei Noah's Ark Lab

Habib Hajimolahoseini

Huawei Toronto Research Centre

Sarath Chandar

Polytechnique Montreal, MILA

Lukas Cavigelli

Huawei Zurich Research Center

Pascal Poupart

University of Waterloo

Di Niu

University of Alberta

https://eiw2024.github.io/

# CLOSING REMARKS 16:30-16:45

Prompt:

People watching a presentation about artificial intelligence while being sad that the conference is ending in Vancouver

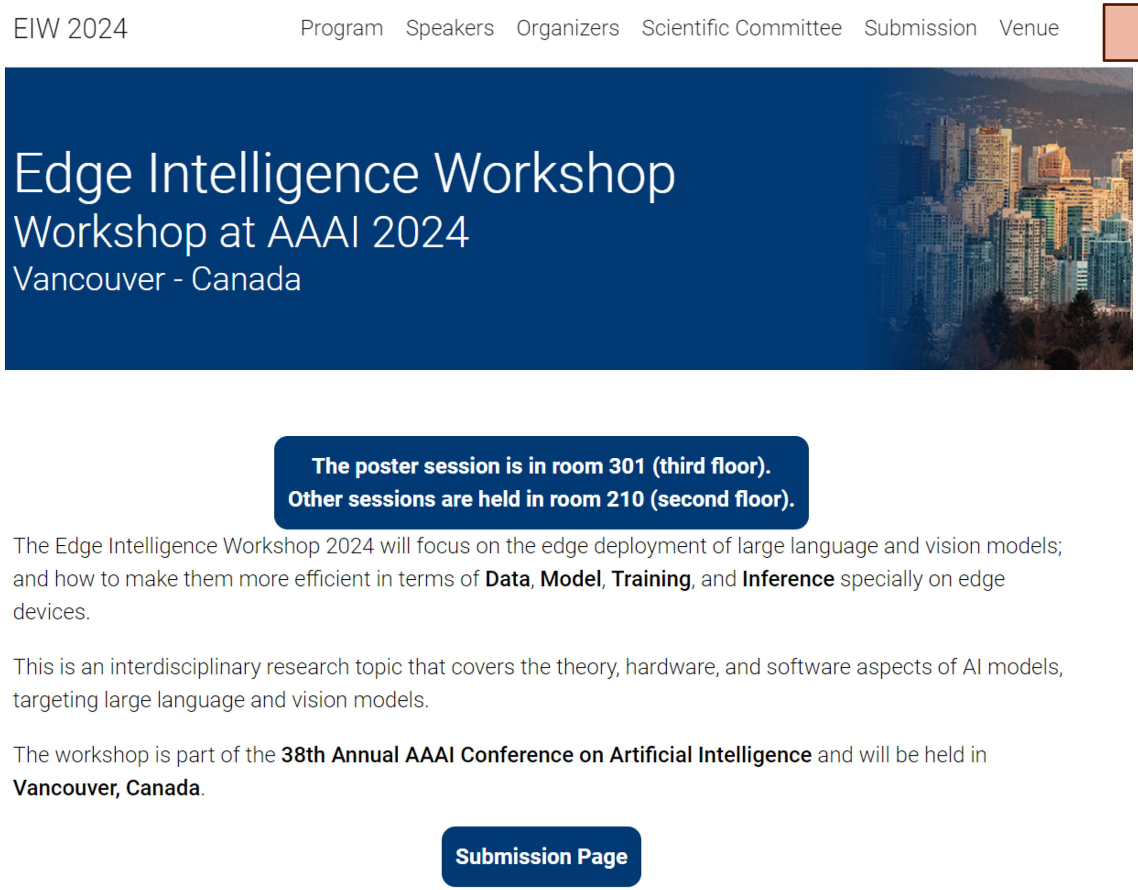
# EDGE INTELLIGENCE CLOSING

AAAI-EIW-III

# ACCEPTED PAPERS

https://eiw2024.github.io/

Accepted papers



Edge Intelligence Workshop
Workshop at AAAI 2024
Vancouver - Canada

**The poster session is in room 301 (third floor).**
**Other sessions are held in room 210 (second floor).**

The Edge Intelligence Workshop 2024 will focus on the edge deployment of large language and vision models; and how to make them more efficient in terms of **Data**, **Model**, **Training**, and **Inference** specially on edge devices.

This is an interdisciplinary research topic that covers the theory, hardware, and software aspects of AI models, targeting large language and vision models.

The workshop is part of the **38th Annual AAAI Conference on Artificial Intelligence** and will be held in **Vancouver, Canada**.

https://eiw2024.github.io/

# ORGANIZERS

Warren Gross

McGill University

Vahid Partovi Nia

Polytechnique Montreal, Huawei Noah's Ark Lab

Andrea Lodi

Cornell Tech

Shahrokh Valaee

University of Toronto

Melika Payvand

UZH and ETH Zurich

Mehdi Rezagholizadeh

Huawei Noah's Ark Lab

Habib Hajimolahoseini

Huawei Toronto Research Centre

Mouloud Belbahri

Layer6AI TDBank

Mohammadreza Tayaranian

McGill University

Yuanhao Yu

Huawei Noah's Ark Lab

Ali Edalati

Huawei Noah's Ark Lab

# KEYNOTE SPEAKERS

Diana Marculescu

University of Texas at Austin

Sarath Chandar

Polytechnique Montreal, MILA

Lukas Cavigelli

Huawei Zurich Research Center

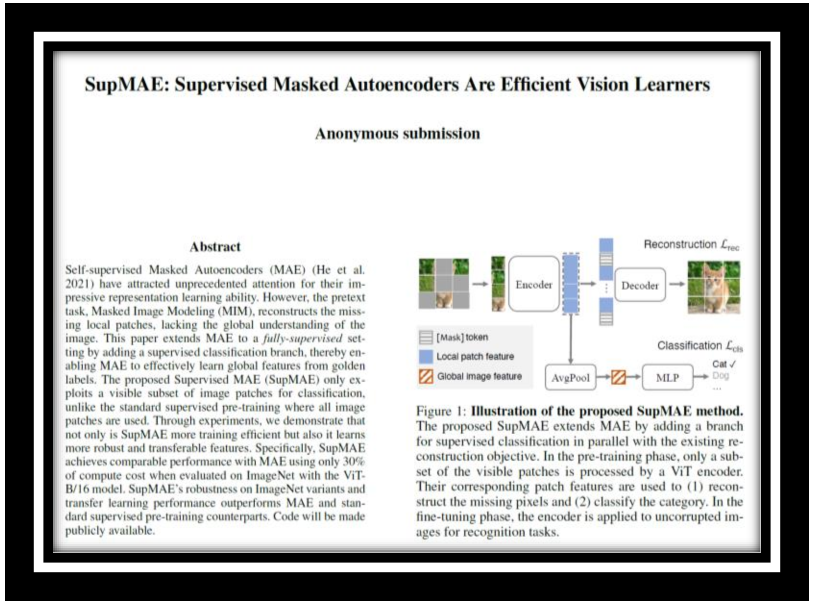
Pascal Poupart

University of Waterloo

Di Niu

University of Alberta

# NOMINATED PAPERS



SupMAE: Supervised Masked Autoencoders Are Efficient Vision Learners

Anonymous submission

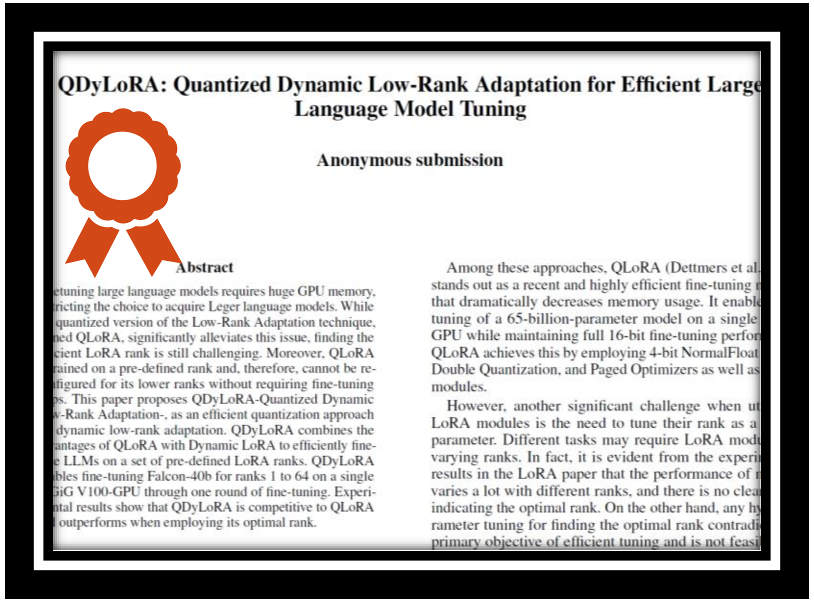
Abstract

Self-supervised Masked Autoencoders (MAE) (He et al. 2021) have attracted unprecedented attention for their impressive representation learning ability. However, the pretext task, Masked Image Modeling (MIM), reconstructs the missing local patches, lacking the global understanding of the image. This paper extends MAE to a *fully-supervised* setting by adding a supervised classification branch, thereby enabling MAE to effectively learn global features from golden labels. The proposed Supervised MAE (SupMAE) only exploits a visible subset of image patches for classification, unlike the standard supervised pre-training where all image patches are used. Through experiments, we demonstrate that not only is SupMAE more training efficient but also it learns more robust and transferable features. Specifically, SupMAE achieves comparable performance with MAE using only 30% of compute cost when evaluated on ImageNet with the ViT-B/16 model. SupMAE's robustness on ImageNet variants and transfer learning performance outperforms MAE and standard supervised pre-training counterparts. Code will be made publicly available.

Figure 1: **Illustration of the proposed SupMAE method.** The proposed SupMAE extends MAE by adding a branch for supervised classification in parallel with the existing reconstruction objective. In the pre-training phase, only a subset of the visible patches is processed by a ViT encoder. Their corresponding patch features are used to (1) reconstruct the missing pixels and (2) classify the category. In the fine-tuning phase, the encoder is applied to uncorrupted images for recognition tasks.

## SupMAE

- Feng Liang (The University of Texas at Austin)
- Yangguang Li (SenseTime Group Limited)
- Diana Marculescu (The University of Texas at Austin)



QDyLoRA: Quantized Dynamic Low-Rank Adaptation for Efficient Large
Language Model Tuning

Anonymous submission

Abstract

etuning large language models requires huge GPU memory,
ricting the choice to acquire Leger language models. While
quantized version of the Low-Rank Adaptation technique,
ned QLoRA, significantly alleviates this issue, finding the
cient LoRA rank is still challenging. Moreover, QLoRA
rained on a pre-defined rank and, therefore, cannot be re-
figured for its lower ranks without requiring fine-tuning
ps. This paper proposes QDyLoRA-Quantized Dynamic
v-Rank Adaptation-, as an efficient quantization approach
dynamic low-rank adaptation. QDyLoRA combines the
antages of QLoRA with Dynamic LoRA to efficiently fine-
e LLMs on a set of pre-defined LoRA ranks. QDyLoRA
bles fine-tuning Falcon-40b for ranks 1 to 64 on a single
GiG V100-GPU through one round of fine-tuning. Experi-
ntal results show that QDyLoRA is competitive to QLoRA
outperforms when employing its optimal rank.

Among these approaches, QLoRA (Dettmers et al.
stands out as a recent and highly efficient fine-tuning
that dramatically decreases memory usage. It enabl
tuning of a 65-billion-parameter model on a single
GPU while maintaining full 16-bit fine-tuning perfor
QLoRA achieves this by employing 4-bit NormalFloat
Double Quantization, and Paged Optimizers as well as
modules.

However, another significant challenge when u
LoRA modules is the need to tune their rank as a
parameter. Different tasks may require LoRA mod
varying ranks. In fact, it is evident from the experi
results in the LoRA paper that the performance of
varies a lot with different ranks, and there is no clea
indicating the optimal rank. On the other hand, any h
rameter tuning for finding the optimal rank contradi
primary objective of efficient tuning and is not feasi

## QDyLoRA

- Hossein Rajabzadeh (University of Waterloo)
- Mojtaba Valipour (University of Waterloo)
- Marzieh Tahaei (Huawei Noah's Ark Lab)
- Hyock Ju Kwon (University of Waterloo)
- Ali Ghodsi (University of Waterloo)
- Boxing Chen (Huawei Noah's Ark Lab)
- Mehdi Rezagholizadeh (Huawei Noah's Ark Lab)

# NEXT EIW 2026 (BE INVOLVED WITH US)

Prompt:

edge intelligence workshop happening in an unknown location in 2026

https://eiw2024.github.io/

# SEE YOU IN TWO YEARS EIW 2026

https://eiw2024.github.io/