International Journal of Computer Vision (2024) 132:1009–1011
https://doi.org/10.1007/s11263-023-01941-4

EDITORIAL

Guest Editorial: Special Issue on the Promises and Dangers of Large Vision Models

Kaiyang Zhou[1] · Ziwei Liu[2] · Xiaohua Zhai[3] · Chunyuan Li[4] · Kate Saenko[5]

Accepted: 19 October 2023 / Published online: 1 November 2023

1 Introduction

This special edition of the International Journal of Computer Vision (IJCV) presents a curated selection of papers that offer a comprehensive overview of the most recent advancements in the realm of large vision models (LVMs). In recent years, LVMs have demonstrated their immense potential across a wide spectrum of computer vision applications. These applications span from general computer vision challenges like image classification and generation, to highly specialized tasks such as autonomous driving and robotics. While the outcomes have been promising, several key challenges pertaining to LVMs have yet to be fully addressed. These include questions about how to enhance the efficiency of pre-training, particularly in terms of data utilization, how to effectively transfer pre-trained models to downstream tasks with limited labeled examples, and the robustness of LVMs for deployment in real-world applications, among others.

✉ Kaiyang Zhou
kyzhou@hkbu.edu.hk

Ziwei Liu
ziwei.liu@ntu.edu.sg

Xiaohua Zhai
xzhai@google.com

Chunyuan Li
chunyl@microsoft.com

Kate Saenko
saenko@bu.edu

[1] Hong Kong Baptist University, Hong Kong SAR, China

[2] Nanyang Technological University, Singapore, Singapore

[3] Google Brain, Zurich, Switzerland

[4] Microsoft Research, Redmond, US

[5] Boston University, Boston, US

This special issue garnered a total of 24 submissions. After undergoing rigorous peer review in line with the journal's high standards, 11 papers have been accepted, resulting in an acceptance rate of 45.8%. The accepted papers showcase cutting-edge advancements in the domain of LVMs and can broadly be categorized into three groups: pre-training, transfer learning, and robustness.

2 Pre-training of LVMs

Pre-training has become a common practice in real-world applications involving large neural networks. During the pre-training phase, neural networks that are initialized randomly are typically trained using extensive datasets of unlabeled images and learning objectives that are agnostic to specific tasks. One of the most successful and widely adopted pre-training methods is the masked autoencoder (MAE) technique, which is founded on the concept of filling in missing image patches. In addition to reconstructing the pixels of masked images, Chen et al. introduce an auxiliary objective function aimed at enhancing representation learning. This auxiliary function focuses on predicting the features of the masked image patches. Despite its simple design, this approach yields significant improvements across a broad spectrum of downstream tasks, including object detection and semantic/instance segmentation.

Another perspective on enhancing the MAE approach is presented by Gao et al. In their approach, they employ an off-the-shelf teacher model to regularize the encoder's output features, thereby minimizing the discrepancy between the teacher's and the encoder's features. The teacher-student learning strategy employed by Gao et al. not only enhances the performance of transfer learning in downstream tasks but also expedites the training process.

3 Transfer Learning of LVMs

A crucial step following pre-training is transfer learning, where a model initially trained on a large dataset undergoes fine-tuning on a different yet related downstream dataset. One significant challenge in applying transfer learning to LVMs lies in achieving parameter efficiency. LVMs are often excessively large for fine-tuning on datasets of moderate sizes. In this special issue, Zhao et al. tackle this challenge with Salient Channel Tuning (SCT), a parameter-efficient fine-tuning method that selectively updates only a small subset of feature channels, approximately three orders of magnitude smaller than the original pre-trained model. These channels are identified as being essential for a specific downstream task through the application of an L2 norm to feature activations within each channel, and channels with the highest values are chosen for tuning.

Zhang et al., on the other hand, address the issue of efficient transfer learning using the encoder-decoder Transformer architecture for semantic segmentation. In this approach, a condensed architecture based on down-sampling query tokens is introduced on the encoder side. On the decoder side, a lightweight attention-to-mask (ATM) module is devised. This module employs learnable class tokens as queries to identify spatial locations that exhibit high compatibility with each class. The resulting architecture not only proves to be robust in standard semantic segmentation scenarios but also demonstrates applicability to the challenge of continual learning, where only the ATM module is updated for each new task.

Lastly, Wu et al. undertake a comprehensive benchmark of multiple classifier learning methods, including linear probing and text feature transfer, and conduct a thorough analysis using a wide array of vision datasets—ranging from 2D images to videos and 3D point clouds. Their findings suggest that the most robust method is the transfer of text features generated by a text encoder.

In addition to addressing parameter efficiency, data efficiency emerges as a significant concern in the realm of downstream transfer learning, particularly when only a limited number of labelled samples are available for training. Within this special issue, several research studies grapple with the challenge of data efficiency for large vision-language models. These models, in contrast to unimodal LVMs, entail an additional neural network that learns to map text onto a shared embedding space.

One widely used approach for adapting vision-language models to low-data scenarios is soft prompt learning, which involves updating a small subset of continuous vectors within the input space of the text neural network. Bulat and Tzimiropoulos enhance the generalization capability of soft prompt learning by introducing a regularization term that minimizes the disparity between learnable prompts and manually crafted prompts.

Meanwhile, Wang et al. tackle the problem of few-shot action recognition by adapting vision-language models through a two-module approach. The first module employs video-text contrastive learning, which formulates videos and their corresponding text features in a contrasting framework. The second module utilizes text features to refine visual prototypes, facilitating improved performance in few-shot action recognition.

4 Robustness of LVMs

Robustness holds paramount importance when it comes to the practical deployment of LVMs in real-world scenarios. Within this special issue, robustness is explored from four distinct perspectives: spurious correlations, handling out-of-distribution (OOD) examples, addressing task bias, and dealing with imbalanced data.

Ghosal et al. embark on an extensive investigation into the performance of large Vision Transformers on datasets containing spurious correlations. Their findings reveal that enhancing model robustness to spurious correlations can be achieved by incorporating more training data and deploying larger models. The study also delves into the impact of various factors, such as extended tuning time, data imbalances, and low-shot learning.

Ming and Li concentrate their efforts on large vision-language models and conduct a comprehensive assessment of different adaptation methods' influence on OOD detection in a few-shot context. The results underscore the potential of adaptation methods to enhance OOD detection, particularly when paired with an appropriate OOD score function, with prompt learning emerging as the preferred choice.

Similarly, in the domain of vision-language models, Menon et al. unveil the presence of task bias within representations acquired through contrastive language-image pre-training. To mitigate this issue, the authors draw inspiration from the concept of visual prompting and propose the introduction of an object guidance token into the model's input image space.

Wang et al. delve into the challenge of adapting vision-language models to datasets characterized by imbalanced data distributions. Their study offers a comprehensive benchmark that encompasses various imbalanced learning algorithms, such as class-balanced loss functions and state-of-the-art two-stage methods.

5 Summary

In summary, the 11 contributions featured in this special issue provide a wide array of perspectives for addressing some of the most pressing challenges in the domains of large vision and vision-language models. This special issue will stimulate the interest of both seasoned experts in the field and those seeking an up-to-date overview of the cutting-edge research in foundation models. We firmly believe that these papers have the potential to make a significant impact not only within the computer vision community but potentially extending their influence beyond it.

We extend our heartfelt gratitude to the dedicated reviewers who invested their valuable time and effort in meticulously assessing the papers and offering constructive feedback to the authors. Additionally, we wish to express our appreciation to the diligent editorial team at Springer, whose invaluable assistance was instrumental in bringing this special issue to fruition.