Image Coding for Machines with Object Region Learning

Takahiro Shindo, Taiju Watanabe, Kein Yamada, Hiroshi Watanabe

Graduate School of Fundamental Science and Engineering, Waseda University, Tokyo, Japan
taka_s0265@ruri.waseda.jp, lvpurin@fuji.waseda.jp, stslm738.ymd@toki.waseda.jp, hiroshi.watanabe@waseda.jp

2024 IEEE 21st Consumer Communications & Networking Conference (CCNC) | 979-8-3503-0457-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CCNC51664.2024.10454864

***Abstract*—Compression technology is essential for efficient image transmission and storage. With the rapid advances in deep learning, images are beginning to be used for image recognition as well as for human vision. For this reason, research has been conducted on image coding for image recognition, and this field is called Image Coding for Machines (ICM). We propose an image compression model that learns object regions. This compression model can cleanly decode only the object regions in the image. In the experiments, we verify the effectiveness of our model by comparing it with previous methods.**

***Index Terms*—Image coding for Machines, ICM, image compression, object detection, segmentation**

I. Introduction

Image compression is an important technology for handling a lot of images and videos. For this reason, image coding methods such as JPEG [1], and VVC [2] have been created. These image compression methods are composed of hand-crafted algorithms. Neural network based image compression (NIC) has also been the subject of much research in recent years. Many NIC models beyond VVC, have been proposed and are expected to be widely used in the future.

With the development of image recognition technology, opportunities to use techniques such as object detection and segmentation are increasing. Hence, it is necessary to devise an ICM method. We propose an NIC model which learns the object region in the images. Our proposed model is trained using Object-MSE-loss. The Object-MSE-loss is the difference between the object region in the input and output image. By applying this loss, only the object regions is decoded cleanly. In the experiments, we compare our model with the SOTA compression model for human vision [3] and ICM method [4] to demonstrate the effectiveness of the proposal.

II. Related Works

In recent years, there has been a lot of research on NIC. These models targeting human vision are trained using the following loss functions:

$$\mathcal{L}_h = \mathcal{R}(y) + \lambda \cdot MSE(x, \hat{x}). \tag{1}$$

In (1), y is the encoder output of the NIC model, $\mathcal{R}(y)$ is the bitrate of y. x represents the input image, and $\hat{x}$ represents the reconstructed image. MSE represents the mean squared error function and λ is a constant to control the bitrate.

Unlike these models, many methods in ICM use task-loss to train the NIC model. The coded image is input to the recognition model to calculate the task-loss. However, recognition model is black box, and training NIC models using task-loss is difficult. Therefore, in general, the NIC model is trained using a loss function that adds task-loss to (1). In this case, the loss function can be expressed as follows:

$$\mathcal{L}_m = \mathcal{R}(y) + \lambda_1 \cdot MSE(x, \hat{x}) + \lambda_2 \cdot \mathcal{M}(\hat{x}). \tag{2}$$

In (2), $\mathcal{R}$, MSE, y, x, and $\hat{x}$ have same meaning in (1). $\mathcal{M}(\hat{x})$ is task-loss, and λ_1 and λ_2 are constants to control bitrate.

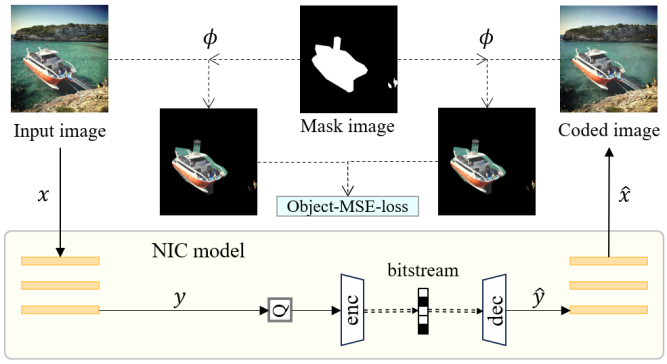

Fig. 1. The proposed training method of the NIC model.

III. Proposed Method

The problem with the task-loss-based approach is that the NIC model is difficult to train. Considering this problem, we propose an NIC model that learns the object region. The Object-MSE-loss is used to train the NIC model. This loss represents the MSE-loss of the object region in the image. By training the NIC model with the Object-MSE-loss, we can create a compression model that only decodes the object region in the image. The proposed training method of the NIC model is shown in Fig. 1. Object-MSE-loss is expressed as follows:

$$Object\_MSE(a, b) = MSE(a \odot m_a, b \odot m_b). \tag{3}$$

In (3), a and b represent certain images, and m_a and m_b are the binary masks corresponding to a and b respectively. Using (3), the proposed loss function is expressed as follows:

$$\mathcal{L}_p = \mathcal{R}(y) + \lambda \cdot Object\_MSE(x, \hat{x}). \tag{4}$$

In (4), $\mathcal{R}$, y, x, $\hat{x}$, and λ have the same meaning in (1).

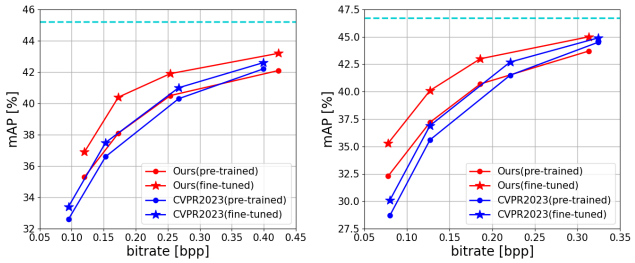

Fig. 2. Compression performance in object detection accuracy of YOLOv5. The left figure shows compression performance for COCO, and the right figure shows the same for VisDrone.

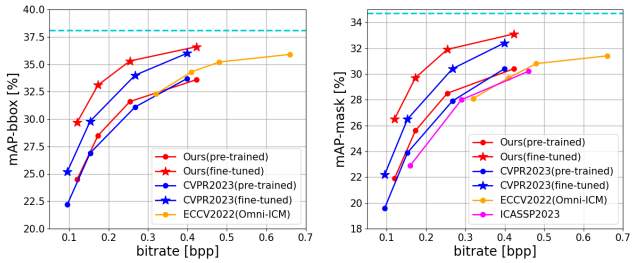

Fig. 3. Compression performance in image recognition accuracy of Mask-RCNN. The left and right figures show the compression performance in detection accuracy and instance segmentation accuracy, respectively.

Fig. 4. Examples of coded images of the COCO2017 dataset with the proposed NIC model. The upper and lower rows are the input and output coded images, respectively.

IV. Experiments

A. Evaluation Method

We modify the loss function used to train the NIC model proposed by J. Liu *et al* [3]. This NIC model is trained using (1) as the loss function. We train this model using (4). The training dataset is COCO2017 [6]. Segmentation map of this dataset is used to create a mask image for calculating the Object-MSE-loss. We trained NIC models using COCO-Train, and four different λ (0.05, 0.02, 0.01, 0.005) are used in (4). As for a comparison, NIC models trained with (1) are prepared. To ensure a fair comparison, COCO-Train is used for training, and four λ (0.01, 0.005, 0.002, 0.001) are used.

We investigate the image compression performance of these NIC models in image recognition. YOLOv5 [7] and Mask-RCNN [8] are used as image recognition models. First, we encode COCO2017 and VisDrone [9] validation datasets with these NIC models. Coded images produced by the proposed method are shown in Fig. 4. The coded images of the proposed NIC model are clean in the object region and unclear in the non-object region. These encoded images are input to pre-trained YOLOv5 and Mask-RCNN to measure the image recognition accuracy. Next, the COCO-Train dataset is also encoded with these NIC models, and these images are used to fine-tune YOLOv5 and Mask-RCNN. The image recognition accuracy of these fine-tuned models is then measured.

B. Results

The results of image recognition are shown in Fig. 2 and Fig. 3. The light blue dotted line indicates the image recognition accuracy of the uncompressed images. The red line shows the coding performance of the proposed method, and the blue line shows that of the comparison. The orange and pink lines indicate the coding performance of the method proposed by R. Feng *et al* [4] and B. Li *et al* [5], respectively. In both cases, the best coding performances are obtained when the proposed NIC model is used for the compression and the fine-tuned model is used for the image recognition.

V. Conclusion

We proposed an NIC model which learns the object region in images. Training the NIC model with object-MSE-loss, we created a model that decodes object regions of images intensively. Experimental results show that the proposed model is effective as ICM method. Future work is required to improve the coding performance by reducing the texture of images.

Acknowledgment

These research results were obtained from the commissioned research (No.05101) by National Institute of Information and Communications Technology (NICT), Japan.

References

[1] G. K. Wallace, “The JPEG still picture compression standard,” IEEE Transactions on Consumer Electronics, vol. 38, no. 1, Feb. 1992.
[2] Versatile Video Coding, Standard ISO/IEC 23090-3, ISO/IEC JTC 1, Jul. 2020.
[3] J. Liu,*et al*, “Learned Image Compression with Mixed Transformer-CNN Architectures,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14388-14397.
[4] R. Feng*et al*, “Image Coding for Machines with Omnipotent Feature Learning,” Computer Vision - ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13697. 2022, pp 510-528.
[5] B. Li*et al*, ”ROI-Based Deep Image Compression with Swin Transformers,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1-5.
[6] T. Y. Lin*et al*, ”Microsoft COCO: Common Objects in Context,” Computer Vision - ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. 2014, pp 740-755.
[7] G. Jocher *et al*., ”ultralytics/yolov5: v7.0-yolov5 sota realtime instance segmentation,” Zenodo, Nov., 2022.
[8] K. He*et al*, ”Mask r-cnn,” Proceedings of the IEEE international conference on computer vision (ICCV), 2017, pp. 2961-2969.
[9] D.Du *et al*., ”VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 0-0.