2P5-4

Application of hybrid attention transformer for increasing spatial resolution of ultrasound human breast images

Chengyen Wu[1], Chikayoshi Sumi [1†] ([1]Info. Sci., Graduate school of Sophia Univ.)

1. Introduction

Recently, Transformers have demonstrated remarkable success in image super-resolution. However, the performance of traditional Transformers is limited by using relatively small amounts of input information. If the input information could be increased, the effectiveness of super-resolution must be further improved. Based on this idea, the Hybrid Attention Transformer (HAT) was developed[1)] and has achieved promising results. Compared to simpler models such as Deep Denoising Super Resolution CNN (DDSRCNN)[2)], HAT has much more complex structures and is more effective at utilizing input information. In this report, we classified ultrasound echo human breast image learning data into four categories—benign, malignant, normal, and all; then HAT and DDSRCNN were respectively trained on these categories and for comparison, the results were evaluated visually and quantitatively using the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) across different categories.

2. Methods

2.1 HAT

HAT is a transformer-based super-resolution model that outperforms traditional transformer-based models.[1)] The improvement is achieved by combining channel attention and window-based self-attention, which enhances the utilization of input pixel information. Additionally, HAT incorporates overlapping cross-attention modules that strengthen the interaction between neighboring window features, leading to more accurate and detailed image reconstruction.

2.2 DDSRCNN

Used for comparison is DDSRCNN. DDSRCNN[2] has convolutional and deconvolutional layers, which reduces noises while simultaneously achieving a high spatial resolution.

2.3 Human breast data

Totally, 780 human breast ultrasound images[3)] were used (500 × 500 pixel average, 1 to 5 MHz), which were comprised of 437 benign, 210 malignant, and 133 normal images. To make the ground truth (GT) data, the images were resized to 256 × 256 pixels at first. Next, the high-resolution (HR) images were down-sampled to 128 × 128 pixels to make low-resolution (LR) image input data for HAT; and the LR data were subsequently up-sampled to 256 × 256 pixels to make LR input data for DDSRCNN. The data were classified into 4 datasets: benign (LRb), malignant (LRm), normal (LRn) and all images (LRa), each of which was trained and tested on each other. **Fig. 1** shows the examples of LRb, LRm and LRn, along with their corresponding PSNR values. Each dataset was used in a ratio of learning, 7.0: evaluation, 1.5: testing, 1.5. The hyperparameters were for HAT, the learning rate, 0.0002 and the number of epochs, 70; and for DDSRCNN, the learning rate, 0.0001, the initial number of epochs, 300, weight decay, 0.01, and we set up a counter with an initial value of 10. If the Mean Squared Error (MSE) of the validation data does not decrease during the epoch, the counter decreases by 1 until it reaches 0; if the MSE decreases, the counter is reset to 10. This mechanism prevented overfitting.

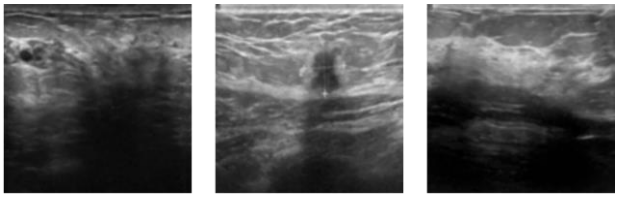

Fig. 1 Examples of LR input images (left) LRb, (middle) LRm, and (right) LRn.

3. Results

Fig. 2 to **Fig. 5** respectively show for the benign, malignant, normal, and mixed learning models of (upper) HAT and (lower) DDSRCNN the examples of resultant images and PSNRs obtained for input data of (left) LRb, (center) LRm and (right) LRn. As shown, it can be visually confirmed that the spatial resolution increases for both models.

Next, the mean PSNR values of all models and the differences between HAT and DDSRCNN are summarized in **Table I**. Regardless the training data,

E-mail: [†]c-sumi@sophia.ac.jp

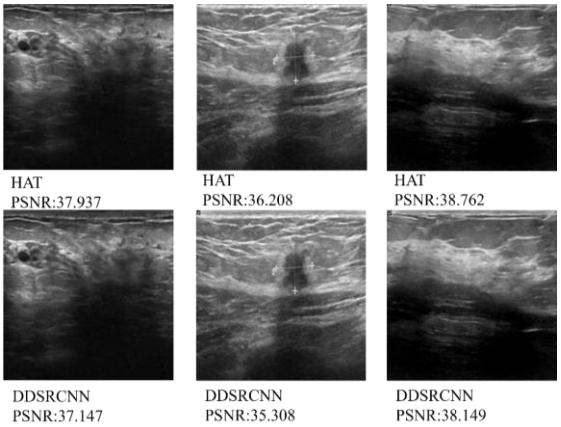

Fig. 2 Images obtained with Benigne (top) HAT and (bottom) DDSRCNN models with different test inputs (left) LRb, (center) LRm, and (right) LRn.

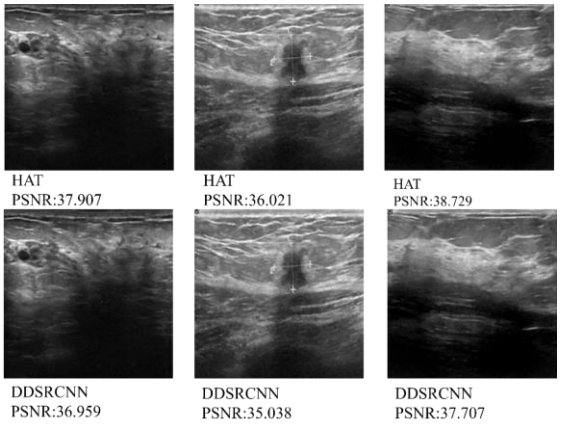

Fig. 3 Images obtained with Malignant models. Specifically, see caption of Fig. 2.

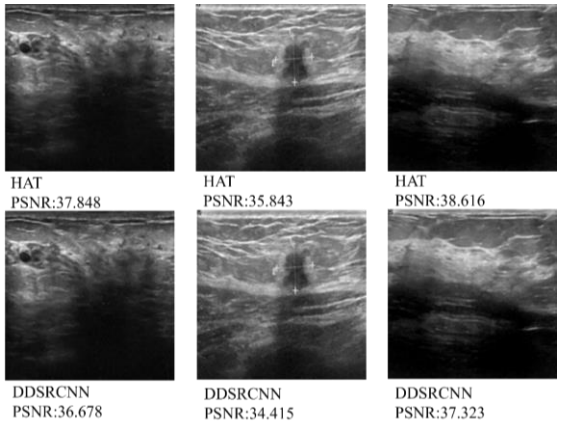

Fig. 4 Images obtained with Normal models. See Fig. 2.

HAT consistently outperforms DDSRCNN from 0.66 to 1.47 dB. Although some individual image results do not follow the expected order, averagely both HAT and DDSRCNN exhibit the more improvement in performance as the amount or diversity of training data is the larger, i.e., the order, Mixed > (≅) Benign > Malignant > Normal model. This was observed for all the inputs except for HAT with LRm, i.e., the Mixed and Benign models were inverted slightly. However, for all the inputs, the order of PSNR differences between HAT and

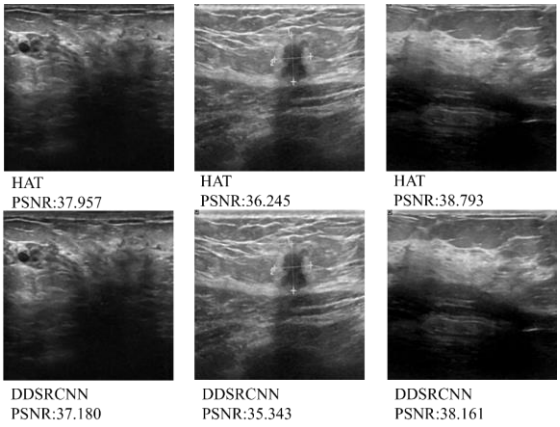

Fig. 5 Images obtained with Mixed model. See Fig. 2.

Table I Mean PSNR values [dB] of all models with respect to every input and differences between HAT and DDSRCNN in parentheses.

HAT/DDSRCNN	LRb	LRm	LRn	LRa
Benigne model	37.56 / 36.68 (0.88)	38.01 / 37.06 (0.95)	37.54 / 36.83 (0.71)	37.69 / 36.82 (0.87)
Malignant model	37.44 / 36.29 (1.15)	37.85 / 36.83 (1.02)	37.52 / 36.67 (0.84)	37.57 / 36.51 (1.06)
Normal model	37.38 / 35.90 (1.47)	37.79 / 36.42 (1.37)	37.52 / 36.40 (1.12)	37.52 / 36.14 (1.38)
Mixed model	37.57 / 36.73 (0.84)	38.00 / 37.19 (0.81)	37.60 / 36.93 (0.66)	37.70 / 36.90 (0.80)

Table II Results about SSIM. Specifically, see Table I.

HAT/DDSRCNN	LRb	LRm	LRn	LRa
Benigne model	0.9687 / 0.9632 (0.0056)	0.9660 / 0.9599 (0.0061)	0.9688 / 0.9635 (0.0052)	0.9680 / 0.9623 (0.0057)
Malignant model	0.9684 / 0.9608 (0.0077)	0.9659 / 0.9588 (0.0070)	0.9686 / 0.9620 (0.0066)	0.9677 / 0.9604 (0.0073)
Normal model	0.9683 / 0.9588 (0.0094)	0.9657 / 0.9568 (0.0090)	0.9686 / 0.9606 (0.0080)	0.9676 / 0.9585 (0.0091)
Mixed model	0.9687 / 0.9626 (0.0062)	0.9661 / 0.9603 (0.0058)	0.9688 / 0.9636 (0.0052)	0.9680 / 0.9621 (0.0059)

DDSRCNN was Normal > Malignant > Benign > Mixed, i.e., the inverse of the order of dataset size. Thus, HAT outperforms DDSRCNN much even when data is less. Moreover, HAT exhibited for all the test data a smaller PSNR range across models than DDSRCNN (differences: 0.19 vs 0.82, 0.21 vs 0.77, 0.08 vs 0.54, and 0.18 vs 0.76). These indicated that HAT has the much higher capability of learning than DDSRCNN and thus, demonstrates the greater stability.

Next, the results of SSIM are similarly summarized in **Table II**, which were similar to those of PSNR outcomes and show that HAT performs better than DDSRCNN.

4. Conclusions

HAT outperforms DDSRCNN in every dataset. Although the the performance of both HAT and DDSRCNN was the better with increasing the number of data or the more diversity of dataset, HAT exhibited the greater stability. To perform the more precise examination, we'll change the amount of training data by increasing/decreasing data, augmentation, etc.

References

1) X. Chen, et al., Proc of IEEE/CVF conf. on computer vision and pattern recognition, 2023, p. 22367.
2) X.-J. Mao, et al., arXiv preprint arXiv:1606.08921 (2016).
3) W. Al-Dhabyani, et al., Data in brief **28** (2020) 104863.