Nested Deformable Multi-head Attention for Facial Image Inpainting

Supplementary Material

Shruti S Phutke and Subrahmanyam Murala
CVPR Lab, Indian Institute of Technology Ropar, Punjab, India

Overview

The Supplementary material includes:

- Difference between existing multi-head attentions and proposed nested multi-head attention.
- The visual results' comparison on the CelebA_HQ dataset corrupted by NVIDIA masks is given in Figure S 2.
- The visual results' comparison on the CelebA_HQ dataset corrupted by QD-IMD masks is given in Figure S 3.
- Computational complexity analysis.
- Ablation Study on Effect of Loss Function

1 Architectural Differences between Proposed NDMAL and Existing Multi-Head Attention

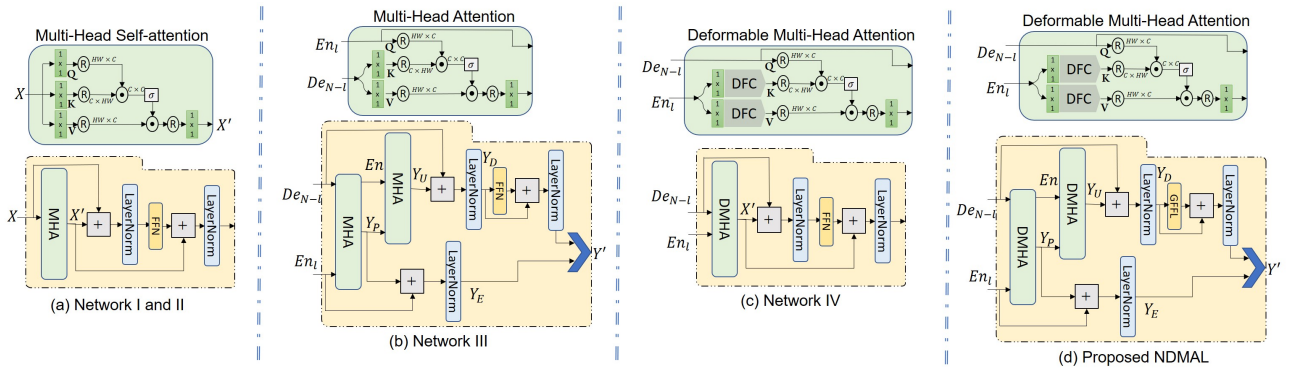

Figure S 1: The architectural difference between proposed nested multi-head attention and existing multi-head attentions (*refer Ablation Study Section in main manuscript*).

In, the existing multi-head self attention the query, key and values are obtained by processing same input with different learnable layers. The output of which is then processed through `Layernorm`→`Feed Forward Network`→`Layernorm` (*see Figure S1 (a).*) Using this attention on either encoder and decoder features will fail to capture the long term dependencies effectively (*see Network I and II in Table 1 and Figure 4 of main manuscript*). Inspired with LUNA [1], we considered the nested attention with encoder and decoder layer to extract the features from both the encoder and decoder layer (see Figure S 1 (b)). This nested multi-head attention fails to extract the valid contextual information (*see Network III in Table 1 and Figure 4 of main manuscript.*) To extract the maximum receptive we added the deformable attention on keys and values. The existing transformer layer is considered with deformable multi-head attention. As, in our proposed network we ought to have lightweight architecture. Utilizing single block with normal attention as shown in Figure S 1 (c), the network is not able to extract the efficient features (*see Network IV in Table 1 and Figure 4 of main manuscript*). The proposed transformer layer consists of both the nested attention and the deformable multi-head attention. This nested deformable multi-head attention helps the network to extract maximum receptive field by capturing the long term dependencies effectively (*see Proposed method in Table 1 and Figure 4 in main manuscript*).

Figure S 2: Qualitative comparison of the proposed method (Ours) with existing methods (GMCNN [3], SN [4], PIC [7], GConv [6], EC [5], RFR [8], HR [9], CTSDG [10], MAT [11]) on CelebA_HQ dataset for NVIDIA [12] mask.

Figure S 3: Qualitative comparison of the proposed method (Ours) with existing methods (EC [5], RFR [8], HR [9], CTSDG [10], MAT [11]) on CelebA_HQ dataset for unknown mask dataset QD-IMD [13].

Table S 1: Computational complexity analysis on $0.01 - 0.6$ mask ratios (NVIDIA [2]) on CelebA-HQ dataset

Method	PSNR	Parameters (Millions)	Run Time	GMAC
GMCNN [3]	24.59	3.115	0.85	-
SN [4]	25.09	54.94	0.12	140.20
EC [5]	26.55	53	0.25	257.96
GConv [6]	25.74	4.05	0.91	111.14
PIC [7]	25.46	3.636	0.98	-
RFR [8]	27.04	31	0.34	412.22
HR [9]	27.36	30	0.22	47.70
CTSDG [10]	27.61	52.14	0.28	61.08
MAT [11]	27.75	60	0.78	435.30
Ours	**28.19**	4.1	**0.08**	15.01

2 Computational Complexity

The computational complexity analysis of proposed method and existing state-of-the-art methods is given in Table S 1. From, Table S 1, it is clear that, our proposed method with lesser complexity gives best results on NVIDIA masks.
Note: We are not able to calculate the GMAC of PIC [55] and GMCNN [42] with the existing source code. Hence we marked "-" for these two methods.

3 Ablation Study on Effect of Loss Function

In this section, we have analysed the effect of various loss functions used to optimize the network while training. The analysis is provided in Table S 2. From this evaluation, we can clearly see that the adversarial loss along with the edge loss and perceptual loss helps the network to work effectively for image inpainting task.

Table S 2: Analysis on effect of loss functions. The evaluation is provided on CelebA-HQ dataset on $0.01 - 0.6$ mask ratio.

Loss Function	PSNR	SSIM	L1	LPIPS	FID
L_1	27.05	0.915	3.896	0.126	8.485
$L_1 + L_{GAN}$	27.76	0.927	3.362	0.108	8.004
$L_1 + L_{GAN} + L_e$	27.92	0.929	3.184	0.091	7.985
$L_1 + L_{GAN} + L_e + L_p$	**28.19**	**0.931**	**2.575**	**0.082**	**6.844**

References

[1] X. Ma, X. Kong, S. Wang, C. Zhou, J. May, H. Ma, and L. Zettlemoyer, "Luna: Linear unified nested attention," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2441–2453, 2021.

[2] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Nvidia irregular mask dataset," in *https://nv-adlr.github.io/publication/partialconv-inpainting*, 2018.

[3] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," *arXiv preprint arXiv:1810.08771*, 2018.

[4] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 1–17.

[5] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 3265–3274.

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480.

[7] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438–1447.

[8] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7760–7768.

[9] G. Wadhwa, A. Dhall, S. Murala, and U. Tariq, "Hyperrealistic image inpainting with hypergraphs," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3912–3921.

[10] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 134–14 143.

[11] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 758–10 768.

[12] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100.

[13] D. Ha and D. Eck, "A neural representation of sketch drawings," *arXiv preprint arXiv:1704.03477*, 2017.