Reconstrução de imagens em super-resolução usando redes neurais convolucionais

Eduardo P. L. Jaqueira, Felippe Durán V. G. Santos, Renato Candido e Magno T. M. Silva

***Resumo*— São propostas duas arquiteturas residuais baseadas em rede neural convolucional para aumentar a resolução de imagens em escala de cinza. Como função custo, foi considerada uma função baseada no índice de similaridade estrutural. Por meio de simulações, verifica-se que as soluções propostas levam a resultados superiores aos obtidos com a interpolação bicúbica.**

***Palavras-Chave*— Super-resolução, rede neural convolucional, interpolação bicúbica, índice de similaridade estrutural.**

***Abstract*— Two residual architectures based on convolutional neural networks are proposed to increase the resolution of grayscale images. We consider a function based on the structural similarity index as a cost function. Through simulations, we observe that the proposed solutions lead to results superior to those obtained with the bicubic interpolation.**

***Keywords*— Super-resolution, convolutional neural network, bicubic interpolation, structural similarity index.**

I. Introdução

A resolução de uma imagem digital determina sua qualidade. Por isso, a reconstrução de imagens em super-resolução a partir de uma única imagem de baixa resolução encontra aplicações em vigilância, imagens médicas, etc. [1]. Dentre os métodos clássicos, destaca-se a interpolação bicúbica, em que os valores dos pixels interpolados são calculados a partir de 16 pixels vizinhos [2]. Apesar de ser uma solução de custo computacional relativamente baixo, as imagens de alta resolução obtidas com essa técnica nem sempre apresentam bordas nítidas e podem conter artefatos. Isso ocorre porque a interpolação bicúbica não introduz nenhuma informação adicional, já que o valor do pixel reconstruído é obtido a partir dos valores dos pixels de sua vizinhança [1].

Para obter melhores resultados, soluções baseadas em aprendizado de máquina têm sido exploradas na literatura, com destaque para a rede neural convolucional (*convolutional neural network* – CNN) (ver [1] e suas referências). A ideia é que se um número suficiente de pares de imagens de baixa e alta resolução for apresentado à CNN, ela pode "aprender" os detalhes inexistentes nas imagens de baixa resolução.

Neste artigo, são propostas duas soluções baseadas em CNN para aumentar a resolução de imagens em escala de cinza. Especificamente, imagens de dimensões $N \times N$ são transformadas em imagens de resolução mais alta com dimensões $2N \times 2N$. Para facilitar o treinamento, são utilizadas conexões residuais entre as camadas do modelo [3]. Além disso, como o erro quadrático médio não é adequado para medir a diferença perceptual entre imagens, considera-se uma função custo baseada no índice de similaridade estrutural (*structural similarity* – SSIM) [4]. Esse índice mede a similaridade entre duas imagens, assumindo valores no intervalo $[-1, 1]$, sendo igual a 1 quando as duas imagens são iguais [4].

Os autores estão com o Depto. de Eng. de Sistemas Eletrônicos, Escola Politécnica da USP, São Paulo, SP, emails: eduardo.jaqueira@usp.br; felippe.santos@usp.br; renatocan@lps.usp.br; magno.silva@usp.br. Este trabalho foi financiado pelo CNPq (121036/2022-7 e 303826/2022-3), FAPESP (2021/02063-6) e CAPES (código de financiamento 001).

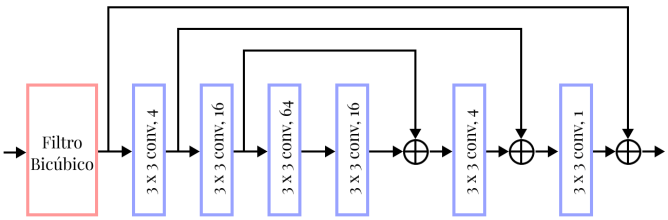
II. Soluções propostas

As arquiteturas propostas estão mostradas na Fig. 1. Na 1ª Abordagem [Fig. 1(a)], aplica-se a interpolação bicúbica na imagem de baixa resolução. A imagem resultante já com as dimensões desejadas entra na CNN composta de seis camadas convolucionais, tendo a ReLU (*rectified linear unit*) como função de ativação. Neste caso, o papel da CNN é melhorar a qualidade da imagem interpolada pelo filtro bicúbico. As dimensões e quantidade dos filtros por camada estão indicadas na figura. Para adequação das dimensões, considera-se o processo *zero padding* antes de cada camada convolucional [5]. As dimensões do tensor de saída da terceira camada ($2N \times 2N \times 64$) possibilitam extrair diferentes características da imagem. Já o tensor de saída da última camada tem $2N \times 2N$, como desejado. Além disso, são utilizadas conexões residuais entre as camadas conforme indicado na Fig. 1(a).

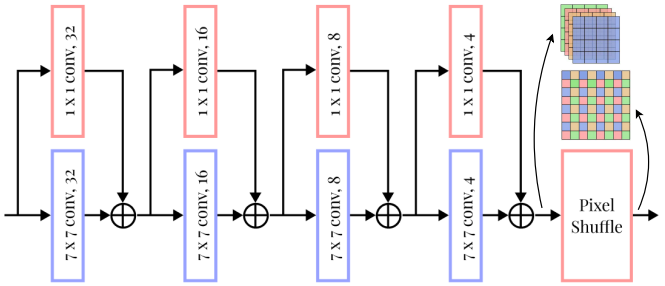
Na 2ª Abordagem [Fig. 1(b)], não se utiliza a interpolação bicúbica. A CNN é formada por quatro camadas convolucionais com ReLU. Neste caso, a rede é responsável por aumentar o tamanho da imagem e ao mesmo tempo melhorar sua qualidade. Em cada camada, duas convoluções são realizadas em paralelo. O tensor de saída dos 32 filtros 7×7 da primeira camada tem dimensões $N \times N \times 32$. Paralelamente nesta camada, o resultado da convolução da imagem com os 32 filtros 1×1 levam a um tensor de mesmas dimensões para que seja possível fazer a conexão residual [3]. Considera-se novamente *zero padding* antes de cada convolução. Esse processo se repete até a quarta camada, cujo tensor de saída tem dimensões $N \times N \times 4$. Esse tensor é então submetido à operação de *pixel shuffle* [6], em que os pixels dos quatro canais são justapostos de modo a gerar uma saída $2N \times 2N$.

III. Simulações

Como banco de dados, foram utilizadas 200 imagens do *Berkeley Segmentation Dataset* [7]. Dessas imagens, 150 foram utilizadas no treinamento e 50 no teste dos modelos. As imagens do conjunto de treinamento foram transformadas em escala de cinza e recortadas considerando $N = 75$. As imagens desejadas com dimensões 150×150 foram subamostradas com um filtro bicúbico para gerar as imagens de baixa resolução com dimensões 75×75. Em ambas as arquiteturas, foram utilizados os mesmos parâmetros de treinamento. Como função custo considerou-se $J = 1 - \text{SSIM}(\mathbf{Y}, \mathbf{D})$, em que $\mathbf{Y}$ é a imagem de alta resolução obtida com o modelo e $\mathbf{D}$ a desejada.



(a) Primeira abordagem



(b) Segunda abordagem

Fig. 1: Arquiteturas das CNN residuais propostas.

Além disso, considerou-se o otimizador Adam com parâmetros $\eta = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ e $\epsilon = 10^{-8}$ [5]. Os pesos e *biases* de uma dada camada foram inicializados de maneira aleatória com distribuição uniforme em $[-1/L, +1/L]$, em que L representa o número de filtros da respectiva camada. O treinamento foi realizado por 10^3 épocas.

Os modelos foram testados com 50 imagens de [7] com $N = 100$ e os resultados estão disponíveis em `https://github.com/Adelkend/SRRNN`. Para ilustrar, apresentam-se os resultados para uma imagem não pertencente à [7] com $N = 170$. O teste apresentado aqui é representativo, de modo que as conclusões obtidas valem para o conjunto de teste. A imagem de alta resolução desejada é mostrada na Fig. 2(a). As imagens de alta resolução obtidas foram comparadas com a desejada por meio do SSIM. A imagem obtida com a interpolação bicúbica é mostrada na Fig. 2(b), apresentando SSIM $= 0{,}877$. As imagens obtidas com as arquiteturas das Figs 1(a) e 1(b) são mostradas nas Figs. 2(c) e 2(d), apresentando respectivamente SSIM de 0,897 e 0,891. Esses valores são obtidos a partir de uma média do SSIM calculado em janelas 7×7. Para analisar as diferenças ao longo da imagem, nas Figs. 2(e) e 2(f) são mostrados o SSIM das imagens das Figs. 2(c) e 2(d), respectivamente.

Observa-se que a qualidade das imagens obtidas com as CNNs propostas é superior à da obtida com interpolação bicúbica, principalmente acima do olho do lagarto. Comparando as imagens das Figs. 2(c) e 2(d), observa-se ainda que a qualidade da imagem obtida com a 2ª Abordagem é ligeiramente superior a da obtida com a 1ª, o que pode ser comprovado pela maior densidade de pixels brancos (SSIM $=$ 1) na Fig. 2(f) em comparação com a Fig. 2(e), fora da pupila. A maior concentração de erros da imagem obtida com a 2ª abordagem está na região da pupila, o que não acontece na 1ª abordagem por causa da interpolação bicúbica. No entanto, esses erros não comprometem a qualidade da imagem e explicam os valores de SSIM muito próximos das duas abordagens. O número de parâmetros treináveis da 2ª Abordagem é cerca de 10 vezes maior que o da 1ª. Isso acaba sendo compensado pelo fato da 2ª Abordagem lidar com imagens menores, enquanto a 1ª já processa imagens com as dimensões desejadas. Consequentemente, elas apresentam aproximadamente o mesmo tempo de processamento.

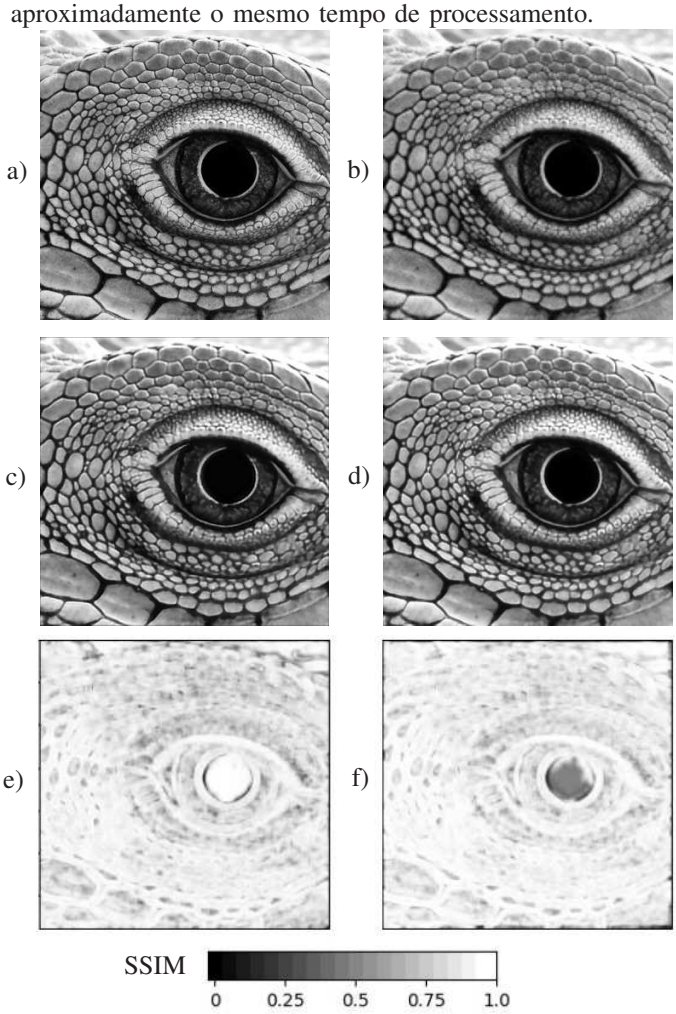


Fig. 2: Imagens de alta resolução (a) Desejada; (b) obtida com a interp. bicúbica, SSIM $= 0{,}877$; (c) obtida com a 1ª Abordagem, SSIM $= 0{,}897$; e (d) obtida com a 2ª Abordagem, SSIM $= 0{,}891$; (e) e (f) SSIM das imagens (c) e (d), respectivamente.

IV. Conclusão

As CNNs residuais propostas foram capazes de obter imagens de alta resolução com qualidade superior à obtida com a interpolação bicúbica. Cabe destacar que a 2ª Abordagem obtém uma solução com qualidade ligeiramente melhor que a da 1ª, tendo aproximadamente o mesmo custo computacional.

Referências

[1] C. Tian *et al.*, "A heterogeneous group CNN for image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[2] R. Keys, "Cubic conv. interpolation for digital image process.", *IEEE Trans. Acoust., Speech, Signal Process.*, v. 29, p. 1153–1160, 1981.

[3] K. He *et al.*, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision Pattern Recognition*, 2016, p. 770–778.

[4] Z. Wang *et al.*, "Image quality assmt.: from error visibility to structural similarity," *IEEE Trans. Image Process.*, v. 13, p. 600–612, 2004.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT, 2016.

[6] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network.", in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016, p. 1874–1883.

[7] D. Martin *et al.*, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics", *8th Int. Conf. Computer Vision*, p. 416–423, 2001.