contecc 2024
CONGRESSO TÉCNICO-CIENTÍFICO DA ENGENHARIA E DA AGRONOMIA

ANÁLISE DA REAL-ESRGAN PARA SUPER-RESOLUÇÃO APLICADA EM IMAGENS DE ANIMES.

LAUAN FERREIRA DE OLIVEIRA[1], OZEIAS BARBOSA DOS SANTOS[2], JOSUE LOPEZ-CABREJOS[3], THUANNE PAIXÃO[4] e ANA BEATRIZ ALAVAREZ[5]

[1]Graduando em Engenharia de Elétrica, UFAC, Rio Branco - AC, lauan.oliveira@sou.ufac.br;
[2]Graduado em Engenharia de Elétrica, UFAC, Rio Branco - AC, ozeiasbss@gmail.com;
[3]Mestrando em Ciências da Computação, UFAC, Rio Branco - AC, josair21@gmail.com;
[4]MSc. Ciências da Computação, UFAC, Rio Branco - AC, thuannepaixao@gmail.com;
[5]Dr. Engenharia Elétrica, UNICAMP, São Paulo - SP, ana.alvarez@ufac.br

Apresentado no
Congresso Técnico Científico da Engenharia e da Agronomia – CONTECC
7 a 10 de outubro de 2024

RESUMO: O processamento de imagens no campo da visão computacional emprega diversas técnicas, a Super-Resolução se destacando como uma das mais significativas. A super-resolução opera na escala do pixel, sendo a unidade fundamental para aprimorar a resolução de imagens. Este estudo foca na implementação da arquitetura Real-ESRGAN, com o objetivo de elevar a resolução de imagens de anime. Um banco de dados contendo imagens de rostos de anime foi selecionado para este propósito. O processo envolveu quatro treinamentos da arquitetura, cada um com um número variado de iterações. A avaliação das imagens geradas pelos modelos em cada treinamento foi realizada qualitativa e quantitativamente, tendo sido esses modelos comparados com o método de interpolação *bicubic*, através de métricas. Os resultados destacaram a notável superioridade da Real-ESRGAN no aprimoramento da resolução de imagens de anime. Este destaque evidencia-se tanto pela análise qualitativa, quanto pela análise quantitativa, reforçando a eficácia da abordagem em relação ao método de interpolação *bicubic*.

PALAVRAS-CHAVE: Processamento de Imagens, Deep Learning, Qualidade Visual, Aplicações de Imagens.

REAL-ESRGAN ANALYSIS FOR SUPER-RESOLUTION APPLIED TO ANIME IMAGES.

ABSTRACT: Image processing in the field of computer vision employs various techniques, with Super-Resolution standing out as one of the most significant. Super-resolution operates at the pixel scale and is the fundamental unit for improving image resolution. This study focuses on the implementation of the Real-ESRGAN architecture, with the aim of increasing the resolution of anime images. A database containing images of anime faces was selected for this purpose. The process involved four trainings of the architecture, each with a varying number of iterations. The images generated by the models in each training session were evaluated qualitatively and quantitatively, and these models were compared with the bicubic interpolation method using metrics. The results highlighted the remarkable superiority of Real-ESRGAN in improving the resolution of anime images. This was evidenced by both qualitative and quantitative analysis, reinforcing the effectiveness of the approach in relation to the bicubic interpolation method.

KEYWORDS: Image Processing, Deep Learning, Visual Quality, Image Applications.

INTRODUÇÃO

Super-Resolução (SR) de imagem é o processo de obtenção de imagens em alta resolução (HR) a partir de imagens de baixa resolução (LR). Trata-se de uma classe importante de técnicas de processamento de imagens no domínio da visão computacional. Tem uma vasta gama de aplicações, como a imagiologia médica (Figueira, 2013), a imagiologia por satélite, a vigilância e segurança e a imagiologia astronômica, (Ledig, 2017) entre outras. Os animes tem grande contribuição cultural e histórica em uma faixa da sociedade, desde seu surgimento no início do século XX, onde há relatos da

existência das primeiras animações no Japão, como afirma (Litten, 2017), o aprimoramento de animações antigas tem um desafio enorme em termos econômicos e de mão de obra. Uma automatização desse processo pode ser realizada utilizando aproximações matemáticas para o aumento de resolução dessas imagens, como são os métodos de aproximação ponderada (Solomon, 2017), os métodos cúbicos ou bicúbicos (Khaledyan, 2020). Posteriormente, foram desenvolvidos métodos mais robustos utilizando redes neurais profundas. Inicialmente as redes neurais convolucionais foram dominantes nesse fim, extraindo características da imagem e melhorando-as por meio de convoluções. No entanto, em um aumento de resolução, a geração de nova informação é importante. Desse modo, as redes generativas adversárias (GANs) demonstraram um papel importante na tarefa de SR em geral. Neste trabalho, aborda-se a tarefa de alcançar uma SR de imagens de anime, levando em conta fatores além da baixa resolução, adicionando ruído e desfoque nas imagens, para não apenas aumentar a resolução da imagem, mas também lidar com problemas comuns nelas, como as degradações mencionadas.

MATERIAL E MÉTODOS

Uma base de dados centrada em imagens de anime foi baixada de (KAGGLE, 2022). Esta base de dados contém 850 imagens em alta resolução, divididas em 760 para o treinamento e 90 imagens de teste para nossa rede neural. Essas imagens foram pré-processadas, aplicando degradações como ruído e desfoque, em primeira e segunda ordem, além da redução da resolução da imagem em uma escala de 4. Isso assegura que a rede neural não apenas realize uma super-resolução de 4 vezes o tamanho da imagem, mas também lide com problemas de degradação nelas.

As GANs são compostos por duas redes neurais onde há a representação de um gerador e um discriminador. O gerador tenta capturar a distribuição de exemplos verdadeiros para a geração de novos exemplos de dados. O discriminador é normalmente um classificador binário, que discrimina os exemplos gerados dos exemplos verdadeiros com a maior precisão possível. Uma rede neural baseada em GAN foi escolhida para SR de imagem. Esta rede foi proposta inicialmente para super-resolução de imagens sintéticas, chamada Real-ESRGAN, que é a terceira atualização das GANs de Super-Resolução, onde foi implementado um discriminador U-Net com normalização espectral para aumentar a capacidade do discriminador e estabilizar a dinâmica de treinamento. Nesta aplicação, pretendemos mudar o enfoque proposto pelo autor (Wang, 2021) para uma abordagem baseada em animes. Dessa forma, foi realizada uma transferência de aprendizado com base no modelo Real-ESRGAN. Este modelo foi treinado 4 vezes com o dataset mencionado anteriormente, modificando em cada treinamento alguns hiperparâmetros, com a finalidade de obter um melhor resultado na aplicação desejada.

O modelo Real-ESRNet pré-treinado com 400.000 iterações e com taxa de aprendizado de 2.10^{-4} foi utilizado seguindo as instruções da documentação. A Real-ESRGAN foi treinada 4 vezes: a primeira com 1.000 iterações e Batch Size de 128, o segundo treinamento foi com 1.000 iterações e Batch Size de 256, o terceiro treinamento da rede ocorreu com 5.000 iterações e Batch Size de 256, e o último treinamento com 40.000 iterações e Batch Size de 256. Esses treinamentos foram realizados utilizando Python 3.11, em uma GPU RTX 3050 e levaram cerca de 2 dias para cada treinamento.

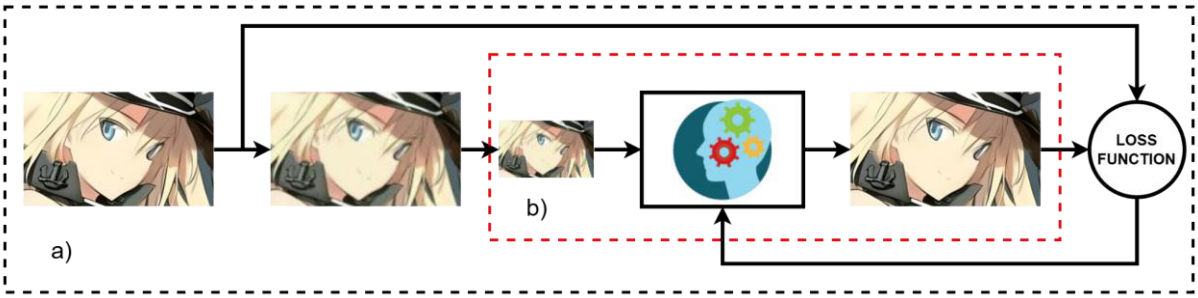
Para comprovar a qualidade dos resultados obtidos, foram utilizadas 4 métricas diferentes que medem a qualidade de uma imagem reconstruída em relação à original, as quais são:

- **Erro Quadrático Médio (MSE)**: É uma medida da qualidade de uma imagem ou sinal que se obtém calculando a média dos erros ao quadrado entre os valores originais e os valores reconstruídos. No contexto de imagens, mede a diferença entre os pixels correspondentes da imagem original e a imagem distorcida ou gerada. Um valor de MSE mais baixo indica uma melhor qualidade de imagem.
- **Razão Pico-Sinal-Para-Ruído (PSNR)**: É uma medida utilizada para avaliar a qualidade de uma imagem ou sinal, comparando a máxima potência possível de um sinal com a potência do ruído que afeta a fidelidade de sua representação. É medida em decibéis (dB). Um PSNR mais alto indica uma melhor qualidade de imagem, onde um valor infinito significa uma reprodução perfeita sem nenhuma distorção.

- **Índice de Similaridade Estrutural (SSIM)**: É uma métrica que avalia a similaridade entre duas imagens, levando em consideração a luminância, o contraste e a estrutura das imagens. Seu valor varia entre 0 e 1, onde 1 indica que as duas imagens são estruturalmente idênticas, e um valor mais próximo de 0 indica menor similaridade.
- **Semelhança Perceptual de Patch de Imagem Apreendida (LPIPS):** É uma métrica que mede a semelhança perceptual entre duas imagens utilizando redes neurais pré-treinadas para esse fim. Ao contrário do MSE e PSNR, o LPIPS captura melhor como os humanos percebem as diferenças entre imagens, considerando aspectos mais complexos da percepção visual. Um valor mais baixo de LPIPS indica uma maior semelhança perceptual entre as imagens.

Na Figura 1, é mostrado o fluxo realizado pela Real-ESRGAN para a SR de imagens, onde a) é o processo de treinamento da rede neural, utilizando um dataset de imagens em alta resolução, aplicando degradações nelas, reduzindo 4 vezes seu tamanho, e alimentando a rede neural com essa imagem pré-processada. A rede neural tenta realizar a reconstrução e SR dessa imagem. Posteriormente, é computada uma função de custo que serve como feedback para ajustar os pesos da rede neural e melhorar seu desempenho iterativamente. E b) representa a aplicação da rede neural uma vez treinada, com uma imagem em baixa resolução alimentando a rede neural, que gera uma imagem em SR a partir da entrada.

Figura 1: Fluxograma do treinamento da Real-ESRGAN.



RESULTADOS E DISCUSSÃO

Os resultados da pesquisa foram analisados a partir da implementação da arquitetura Real-ESRGAN, a qual foi treinada quatro vezes, com 400.000 iterações em cada treinamento, e foram aplicadas as 90 imagens de teste separadas anteriormente do dataset. O hardware utilizado para este experimento foi um notebook Avell com processador Intel i7-12700H, GPU RTX 3050 e 64 GB de memória RAM, utilizando o software Chainner para carregar o modelo treinado e aplicar a SR. Os resultados, são mostrados na Tabela 1, indicando que o treinamento B foi o melhor em 3 das 4 métricas calculadas.

Tabela 1. Resultados dos 4 modelos treinados com o Real-ESRGAN.

Treinamento	MSE	PSNR	SSIM	LPIPS
Treinamento A	0.011	25.48	0.83	0.055
Treinamento B	**0.010**	**26.00**	**0.84**	0.054
Treinamento C	0.018	23.38	0.77	0.068
Treinamento D	0.013	24.69	0.81	**0.041**

Na Tabela 2, são comparados os resultados do nosso melhor treinamento do modelo Real-ESRGAN com os resultados do aumento de resolução pelo método bicúbico, sendo nosso modelo claramente superior ao método tradicional de aumento de resolução, validando esses resultados com todas as métricas de qualidade mencionadas anteriormente.

Tabela 2 Comparação do Real-ESRGAN com o método bicúbico de aprimoramento da resolução.

Método	MSE	PSNR	SSIM	LPIPS
Real-ESRGAN	**0.010**	**2600**	**0.84**	**0.041**
Bicubic	0.014	24.51	0.78	0.208

Os resultados visuais, mostrados na Figura 2, evidenciam uma melhora significativa na imagem em termos de resolução, com notável eliminação de ruído e desfoque da imagem de entrada.

Figura 2: a) Representa a imagem de entrada em uma resolução de 170 x 170 pixels, com ruído gaussiano e desfoque. b) Representa a saída após a aplicação de SR com o modelo Real-ESRGAN.

(a) IN (b) OUT

CONCLUSÃO

O presente trabalho obteve notáveis resultados alcançados por meio da implementação do Real-ESRGAN. Apesar de ter sido realizado um número limitado de iterações, a eficácia do Real-ESRGAN em produzir resultados impressionantes para o aprimoramento de resolução de imagens de anime foi evidenciada, destacando que esses resultados foram obtidos com o gerador pré-treinado. A comparação com abordagens tradicionais de super-resolução enfatizou de forma inequívoca a superioridade dos modelos baseados em deep learning em comparação com métodos convencionais.

AGRADECIMENTOS

Agradeço a Deus, primeiramente, pois não teria chegado até aqui sem sua Graça por mim. Como está escrito na bíblia sagrada: Que diremos, pois, a estas coisas? Se Deus é por nós, quem será contra nós? (Romanos 8:31). Afirmo neste momento que tudo na minha vida depende do Pai Eterno e que sem ele não sou nada.

REFERÊNCIAS

ALLEN-ZHU, Zeyuan; LI, Yuanzhi. Forward super-resolution: How can gans learn hierarchical generative models for real-world distributions. arXiv preprint arXiv:2106.02619, 2021.

ALMAS, Almir. 4K: Estado da Arte. Revista da SET, v. 25, n. 152, p. 26-27, 2015.

CUNHA, João Paulo Zanola. Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. 2019. Tese de Doutorado. Universidade de São Paulo.

FIGUEIRA, Nina M.; OLIVEIRA, Leonardo C. de. Super-resolução: técnicas existentes e possibilidade de emprego às imagens do vant vt-15. Revista Militar de Ciência e Tecnologia, v. 30, p. 3-19, 2013.

HU, Daqi. DragGAN-Based Emotion Image Generation and Analysis for Animated Faces.

KAGGLE. Ganyu- Genshin Impact Anime Faces GAN Training. 2022. Disponível em: https://www.kaggle.com/datasets/andy8744/ganyu-genshin-impact-anime-faces-gan-training. Acesso em: 23 jan. 2024.

LEDIG, Christian et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4681-4690.

SOLOMON, Chris; BRECKON, Toby. Fundamentals of Digital Image Processing: A practical approach with examples in Matlab. John Wiley & Sons, 2011.

VARGAS, Kevin Ian Ruiz. UR-SRGAN: a generative adversarial network for real-world super-resolution with a U-Net-based discriminator. 2022. Dissertação de Mestrado. Universidade Federal de Pernambuco.