Uso de Super Resolução Através de Aprendizado Profundo para Melhora na Performance do OCR em Documentos

Trabalho de Conclusão de Curso

Engenharia de Computação

Gabriel Calazans Duarte de Moura

Orientador: Prof. Byron Leite Dantas Bezerra

Gabriel Calazans Duarte de Moura

Uso de Super Resolução Através de Aprendizado Profundo para Melhora na Performance do OCR em Documentos

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Engenharia de Computação

Escola Politécnica de Pernambuco

Universidade de Pernambuco

Orientador: Prof. Byron Leite Dantas Bezerra

Recife - PE, Brasil

Junho de 2019

Gabriel Calazans Duarte de Moura
Uso de Super Resolução Através de Aprendizado Profundo para Melhora na Performance do OCR em Documentos/ Gabriel Calazans Duarte de Moura. – Recife - PE, Brasil, Junho de 2019-
49 p.
Orientador: Prof. Byron Leite Dantas Bezerra

Trabalho de Conclusão de Curso – Engenharia de Computação
Escola Politécnica de Pernambuco
Universidade de Pernambuco, Junho de 2019.
1.Documentos; 2.Super Resolução; 3.Aprendizagem profunda; 4.OCR

Dedico este trabalho à minha família e amigos.

ESCOLA POLITÉCNICA
DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 02/07/2019, às 11h, reuniu-se para deliberar sobre a defesa da monografia de conclusão de curso do(a) discente **GABRIEL CALAZANS DUARTE DE MOURA**, orientado(a) pelo(a) professor(a) **BYRON LEITE DANTAS BEZERRA**, sob título Uso de Super Resolução Através de Aprendizado Profundo para Melhora na Performance do OCR em Documentos, a banca composta pelos professores:

MÊUSER JORGE SILVA VALENÇA (PRESIDENTE)
BYRON LEITE DANTAS BEZERRA (ORIENTADOR)

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

☒ Aprovada ☐ Aprovada com Restrições* ☐ Reprovada

e foi-lhe atribuída nota: 10,0 (DEZ)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O(A) discente terá 07 dias para entrega da versão final da monografia a contar da data deste documento.

AVALIADOR 1: Prof (a) **MÊUSER JORGE SILVA VALENÇA**

AVALIADOR 2: Prof (a) **BYRON LEITE DANTAS BEZERRA**

AVALIADOR 3: Prof (a)

* Este documento deverá ser encadernado juntamente com a monografia em versão final.

ESCOLA POLITÉCNICA
DE PERNAMBUCO

Autorização de publicação de PFC

Eu, **Gabriel Calazans Duarte de Moura** autor(a) do projeto de final de curso intitulado: **Uso de Super Resolução Através de Aprendizado Profundo para Melhora na Performance do OCR em Documentos;** autorizo a publicação de seu conteúdo na internet nos portais da Escola Politécnica de Pernambuco e Universidade de Pernambuco.

O conteúdo do projeto de final de curso é de responsabilidade do autor.

Gabriel Calazans Duarte de Moura

Gabriel Calazans Duarte de Moura

Orientador(a): **Byron Leite Dantas Bezerra**

Coorientador(a):

Prof, de TCC: **Daniel Augusto Ribeiro Chaves**

02/07/2019

Data: 02/07/2019

Agradecimentos

Em primeiro lugar gostaria de agradecer aos meus familiares pelo apoio que me deram durante essa jornada. Em especial gostaria de agradecer ao meu avô paterno, José Calazans, por ter sido um exemplo que venho tentando seguir há muito tempo.

Também gostaria de agradecer aos meus colegas de trabalho, pelo apoio provido nos últimos meses. Em especial gostaria de agradecer ao Dr. Everton Lacerda, Eduardo Simões e João Paulo Ribeiro pela ajuda e concelhos durante esse período.

Bem como a todos os meus colegas da UPE, em especial a Theo Victor, Camila Luísa, Daniel Vasconcellos, Thaina Paes, Maria Paula, Paula Pithon dentre os vários outros, pela ajuda durante o tempo de graduação.

E gostaria de agradecer ao meu orientador Prof. Dr. Byron Leite, pela orientação durante esse projeto. Bem como aos demais professores pelos ensinos durante os últimos 5 anos.

"We don't get to choose when and where technological progress stops.
We cannot slow down. In fact, we have to speed up.
Our technology excels at removing difficulties and uncertainties from our lives, and so we must seek out ever more difficult, ever more uncertain challenges"
Garry Kasparov[1]

Resumo

A falsificação de documentos é uma prática bem comum na atualidade. Contudo, essas práticas ilegais vem sido contra-atacadas com aplicações de visão computacional tais como o técnicas de reconhecimento óptico de caracteres(OCR, do inglês *Optical character recognition*). Entretanto o OCR possui algumas dificuldades como brilho, inclinação, resolução baixa e a imagem no plano de fundo. Esses erros são bastante recorrentes, por isso o OCR apresenta uma taxa de acerto baixa. Essas aplicações se tornaram mais importantes ainda na era digital, onde várias aplicações têm processos complexos que necessitam cada vez maiores taxas de transferências. Nessas aplicações não há como garantir qualidade desses documentos, e em algumas situações essas aplicações causam deformações, dificultando a taxa de acerto de técnicas de OCRs. Entretanto, várias abordagens têm sido desenvolvidas para suprir essa dificuldade. Imagens de super resolução (SR) é uma dessas técnicas propostas. A SR visa a transcrição de uma imagem em baixa resolução em uma imagem de pseudo-alta resolução, essas técnicas mostraram um grande ganho nos últimos anos com o advento de redes neurais de aprendizagem profunda. Entretanto, não sabíamos o ganho dessas técnicas quando relacionadas com técnicas de OCR em um cenário onde a imagem não é ideal para transcrição devido a outros fatores como a textura do plano de fundo da imagem. Portanto neste trabalho propomos fazer a avaliação do uso de redes de super resolução em um contexto de documento com alto nível de textura em seu plano de fundo, com o objetivo de aumentar a taxa de acerto de programas OCR. Resultados experimentais demonstraram que com nossa abordagem, a taxa de erro por carácter e a taxa de erro por palavra em dois programas populares de OCR, alcançando taxas de 27,76%, mesmo a imagem original em alta definição alcançava 30,10%

Palavras-chave: Documentos; Super Resolução; Aprendizagem profunda; OCR

Abstract

Document falsification has been quite common in nowadays days. However, those illegal practices has been opposed with computing vision applications such as optical character recognition (OCR) techniques. However, OCR's has some issues, mainly with deformation and noise, such as brightness, skew, low resolution and the image background. Those errors are quite common, thus the OCR presents a high accuracy loss. Those applications becomes even more important in the digital age, where many applications have complex processes thus require a high transfer rate. In those applications there is no way to guarantee the quality of these documents, and in some cases those applications tend to deform the information making harder to OCR algorithms to archive higher rates. However, many techniques have been developed to approach some of those deformations. Image super resolution (SR) is one of these techniques that has been proposed. SR aims to transcribes a low resolution image in a fake high resolution image, and have shown a huge improvement in the last years with the advance of deep learning methods. However, we did not know the gain of these techniques when dealing with OCR techniques in a scenario where the image is not ideal for transcription due to other factors such as the texture of background the image. Therefore, in this work we proposed to evaluate the use of super resolution networks in a context of documents with a high texture background to increase the OCR rate. Experimental results show that by using super resolution networks, we were able to present a decrease in the character error rate (CER) and the word error rate (WER) in two popular OCR techniques, by reaching up to 27.76% in the CER error rate, while the original image would only reach 30.10%.

Keywords: Document image; Super Resolution; Deep Learning; OCR

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

SR	Super resolução, do inglês *Super resolution*
SRCNN	*Super Resolution Convolutional Neural Network*
VDSR	*Very deep neural networks for super resolution*
FSRCNN	*Fast Super Resolution Convolutional Network Optical Character Recognition*
DRCN	*Deeply-Recursive Convolutional Network*
RED-Net	*Residual Encoder-Decoder Network*
SRGAN	*Super Resolution Generative Adversarial Network*
DPI	Pontos por polegada , do inglês *Dots per inch*
CNH	Carteira nacional de habilitação
RG	Registro geral
CPF	Cadastro de pessoa física
CER	Taxa de erro por letras, do inglês *Character Error Rate*
WER	Taxa de erro de palavras, do inglês *Word Error Rate*
OCR	Reconhecimento óptico de caracteres, do inglês
PSNR	Relação sinal-ruído de pico, do inglês *peak signal-to-noise ratio*
MOS	Do inglês *Mean opinion score*
GAN	Redes generativas adversárias, do inglês *Generative Adversarial Networks*
ReLU	Do inglês *Rectified linear unit*
PReLU	Do inglês *Parametric rectified linear unit*
MSE	Erro quadrático médio, do inglês *Mean Square Error*
RGB	Do inglês *Red, green blue*PSNR -403

Lista de símbolos

ϕ	Letra grega minuscula phi
$\cap$	Interseção
$\cup$	União

Sumário

1 Introdução

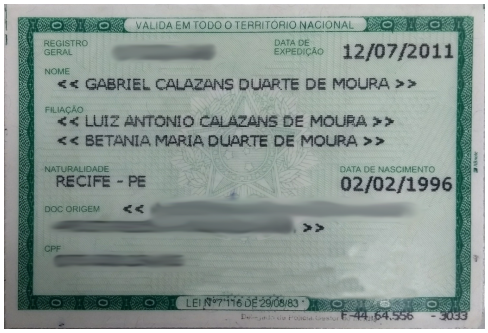
Problemas de super resolução(SR), em especial os casos de super resolução de uma única imagem, vem ganhando a atenção de pesquisadores por décadas. Técnicas de SR focam em reconstruir uma imagem em alta definição a partir de uma de baixa resolução. SR é usada em várias aplicações de visão computacional, como nas àreas de medicina[2], entretenimento[3] e inclusive na área de reconhecimento e extração de textos[4, 5, 6]. Observamos que a super resolução é um problema inverso, já que existem múltiplas soluções em alta definição para qualquer imagem em baixa resolução. Muitos estudos assumem que a imagem de baixa resolução é uma imagem reduzida da sua versão em alta definição, porém outros fatores degradantes como borramento, distorção ou ruído podem ser considerados para aplicações práticas. Para abordar estes problemas, inúmeros métodos de aprendizados vêm sido propostos para aprender a mapear a correlação de pares de imagens em baixa e em alta resolução.

Recentemente, métodos baseados em aprendizado profundo têm conseguido melhoras significativas sobre métodos de SR convencionais. Tais como, SRCNN[7], VDSR[8] e FSRCNN[9]. Tais aplicações vêm mostrando conquistas, devido ao uso de imagens de SR, como mostrado por Lat et al.[5], superando alguns problemas presentes em imagens de baixa resolução, as quais tem menos quantidade de informação. No entanto, é comum que imagens de alta resolução sejam comprimidas em imagens de baixa resolução, para diminuição do tráfico em rede ou diminuição do tamanho da imagem no disco. Devido a estas compressões algumas aplicações, como os programas de reconhecimento ótico de caracteres (OCR do inglês:*optical character recognition*), apresentem uma perda de precisão.

Softwares de OCR são formas de extrair e manipular a informação textual de imagens, contudo os OCRs são extremamente sensíveis à meta informações, tais como erro de foco[10], resolução[4, 5, 6], deformação [11] e ao plano de fundo[12]. Vários algoritmos vêm sido estudados para abordar este problema durante os últimos anos[7, 8, 13, 9, 14, 15, 11, 10]. Uma forma para atenuar esses problemas é o uso de SR. Entretanto, imagens de SR são abordadas na literatura de forma genérica, tentando melhorar a qualidade dos detalhes das imagens. Contudo os resultados das técnicas de SR não foram testadas em casos focados no OCR, onde o plano de fundo contém tantos detalhes. Tais planos de fundo não são incomuns em documentos como mostrado na figura 1.

Esse plano de fundo tem um grande impacto em softwares de OCR, já que a seu principal componente é dependente da binarização, e isso continua sendo um desafio mesmo nos dias atuais[16].

Figura 1 – Exemplo de um documento de identidade brasileiro.



Fonte:Fotografia de autoria própria (2019).

Propomos, neste projeto, avaliar o uso de imagens de SR visando aumentar a taxa de acerto de aplicações de OCR. Foram avaliadas seis diferentes arquiteturas de redes neurais: SRCNN[7], VDSR[8], FSRCNN[9], DRCN[13], RED-Net[14] e SRGAN [15]. Essa avaliação foi feita com o objetivo de medir a eficácia de redes neurais, já existentes, na melhoria das taxas de softwares de OCR. Essas redes foram selecionadas devido a aplicações e resultados anteriores[3, 4, 5, 6]. Também foi considerado o custo computacional dessas redes, devido a limitações físicas de aplicações que utilizam de super resolução.

Durante o treino da rede foram percebidos alguns problemas que levaram a utilizar de uma base privada de imagens em 200 pontos por polegada (DPI do inglês: *dots per inch*). Esta base é formada com 3 tipos diferentes de documentos brasileiros: Carteira nacional de habilitação(CNH), documentos de identidade ou registros gerais(RG) e cadastro de pessoa física (CPF). Estes 3 documentos são os mais comuns no Brasil.

Em nossos testes nós avaliamos as imagens geradas de documentos por redes neurais em 200, 300 e 400 DPI usando dois dos mais populares programas de OCR, o Tesseract OCR[1] e o ABBYY *fine reader*[2]. Essa avaliação demonstrou que é possível diminuir a taxa de erro medida pela taxa de dispositivos de OCR, medidos pela taxa de erro de letras (CER, do inglês *character error rate*) e taxa de erro de palavras(WER, do inglês *word error rate*), em até 25,59%/23,24% quando comparada com a imagem de baixa resolução no cenário do Tesseract, e em 20,19%/17,85% no cenário do ABBYY. Também conseguimos neste projeto uma diminuição da taxa quando comparada com a imagem em alta resolução.

Este trabalho está organizado da seguinte forma: na seção 2 são discutidos os trabalhos relacionados, brevemente o uso da super resolução em documentos, explicando sua história e também revisando alguns trabalhos nesta área. Na seção 3 são apresentados os modelos avaliados, explicando suas arquiteturas e mostrando suas contribuições para as

[1] Tesseract é um projeto de código aberto - https://github.com/tesseract-ocr/tesseract acessado em 23/06/2019

[2] ABBYY é um motor de reconhecimento óptico de caracteres comercial. https://www.abbyy.com/en-eu/ acessado em 23/06/2019

técnicas de SR. Na seção 4 serão explicados os experimentos e a metodologia aplicada, durante este trabalho também serão apresentados os detalhes da implementação. E na seção 5 os resultados serão discutidos.

2 Trabalhos Relacionados

A super resolução vêm sendo alvo de pesquisas desde o seu início da década de 80. Técnicas para aumentar a resolução da imagem vem sendo abordadas de várias formas[17, 18], antes do advento dessas técnicas uma forma popular para aumentar o tamanho das imagens era a interpolação bicúbica[19]. Esta interpolação apresenta algumas vantagens, como um custo computacional pequeno. Entretanto a imagem resultante normalmente apresenta um aspecto borrado. Contudo ainda existem muitas outras técnicas de SR. Por isso, a super resolução pode ser agrupada em duas categorias: a super resolução de múltiplas imagens[20, 21, 22] e a super resolução de única imagem[23, 24, 25].

A super resolução de múltiplas imagens usa de várias imagens, normalmente extraídas de um vídeo ou de fotos sequenciais, da mesma cena para reconstruir os pixeis de uma imagem, assim gerando uma imagem de super resolução. Enquanto, a super resolução de única imagem usa de meta informações, tais como as bordas[18] e informações do domínio da frequência[17], para gerar a imagem de super resolução.

Ambos cenários apresentam resultados interessantes. Entretanto, a super resolução de múltiplas imagens não é viável nesse cenário, devido ao fato que a maioria dos documentos digitalizados são representadas por uma única imagem, e a grande maioria deles não existe mais ou não existe como solicitar ao usuário para tirar mais de uma foto.

A super resolução de uma única imagem se encontra em ascensão desde o trabalho de Dong et al.[7], introduzindo as redes convolucionais de aprendizagem profunda(CNN, do inglês: *Convolutional neural networks*) [26] aos problemas abordados com SR, superando técnicas anteriores.

Essas redes provaram melhorar a relação sinal-ruído de pico networks (PSNR , do inglês *peak signal-to-noise ratio*)[7, 8, 13, 9, 14], superando outras formas de super resolução. Como apresentado em trabalhos anteriores[15], mesmo com as taxas de PSNR melhorando, essa métrica não representa a melhora na qualidade visual da imagem de super resolução, como demonstrado através de outra métrica a *mean opinion score* ou taxa de opinião média (MOS). Portanto, em 2017 foi proposto o uso de redes adversárias generativas(GAN do inglés *Generative Adversarial networks*) [27] por Leding et al.[15] trazendo outro grande avanço para as redes neurais de super resolução.

O impacto decorrente do uso de redes profundas motivou muitos estudos. Uma pesquisa realizada sobre esse conjunto é o uso de redes de aprendizagem profunda para o aumento da taxa de acerto de aplicações de OCR, tais como:

- Em Pamdey et al.[4] é usada uma rede de aprendizagem profunda para extrair as melhores características de métodos de interpolação clássicos, tais como: a interpolação bicúbica, a interpolação por vizinho mais próximo e a interpolação bilinear. A combinação dessas características foi o que eles intitularam como *nonlinear fusion of multiple interpolations* (fusão não linear de múltiplas interpolações). Este trabalho alcançou alguns resultados interessantes na base de documentos binários Tamil, aumentando a taxa de acerto do OCR por 17,89% no melhor cenário.

- Em Lat et al.[5] uma rede generativa adversárias de super resolução é usada, com a imagem no domínio das cores Lab, o treinamento foi feito usando 100 pares de imagens em baixa resolução e em alta resolução, capturadas sobre ambientes controlados. Esse cenário foi avaliado usando o Tesseract 3.02 bem como o ABBYY *fine reader*, tendo um aumento de 21% na taxa de acerto do OCR.

- Em Zhang et al.[6] os autores utilizam redes muito profundas de super resolução (VDSR, do inglês: *Very deep super resolution network*)[8]. Durante seus trabalhos eles propuseram algumas modificações na função de tratamento das bordas e no função de perda. Neste trabalho os autores usaram a base da ICDAR 2015 TextSR. Foi possível alcançar taxas de 78.10% de acerto em OCRs enquanto imagens de alta resolução possuíam taxas de 78.80% OCR de acerto.

Esses trabalhos mostraram bons resultados. Porém, as imagens usadas em suas pesquisas eram, normalmente, binárias ou possuíam um plano de fundo simples (planos de fundo onde o artefato de textura é normalmente pequeno). O que difere do nosso cenário uma vez que as nossas imagens possuem o fundo com alto nível de textura.

3 Modelos Avaliados

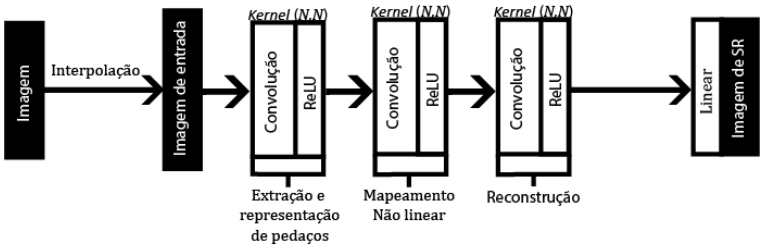
Neste trabalho, foram avaliadas seis redes de SR diferentes, a fim de abordar o problema de resolução do documento. Selecionamos essas redes com base em seu desempenho, custo de computacional e contribuições em trabalhos anteriores [4, 5, 6]. Embora existam muitas outras redes de super resolução, como mostrado por Anwar et al.[28]

3.1 *Super resolution convolutional neural network*

Proposta por Dong et al.[7], a *Super resolution convolutional neural network* (SRCNN) foi a primeira rede neural de aprendizado profundo a abordar ao problema da super-resolução, apresentando resultados superiores às técnicas anteriores[7].

A SRCNN é composta por três operadores: extração e representação de pedaços (do inglês :*patch extraction and representation*), mapeamento não-linear (do inglês :*non-linear mapping*) e a camada de reconstrução (do inglês :*reconstruction layer*) , como mostrado na figura 2. Esses componentes podem ser resumidos como uma camada de convolução seguida por uma ativação de *rectified linear unit* (ReLU), de tal forma que a SRCNN em si não faz a ampliação da imagem, porque a rede foca na redução de ruído da imagem ampliada.

Figura 2 – Arquitetura original da SRCNN.



Fonte:Imagem de autoria própria (2019).

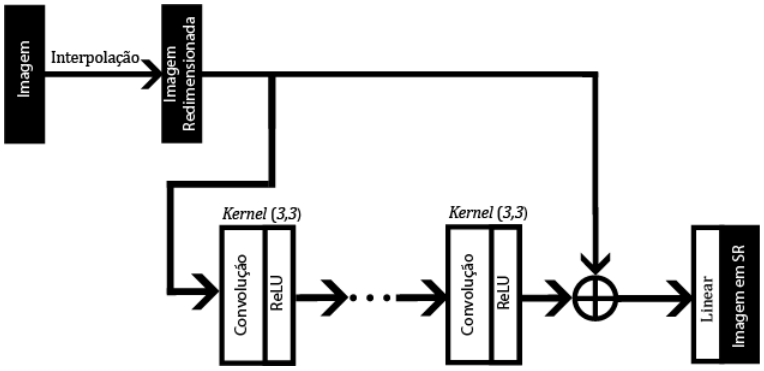
Algumas aplicações foram feitas usando adaptações desta rede neural, como aplicações médicas[2], e também na área de entretenimento, como a Waifu2X [3] que é usado para ampliar a imagem de personagens de animações japonesas.

3.2 *Very deep neural networks for super resolution*

Em 2016, Kim et al. propuseram as *Very deep neural networks for super resolution* (VDSR)[8], baseadas no trabalho do *Visual Geometry Group* [29] e no trabalho de Dong et al.[7]. Este trabalho foi a primeira técnica de rede de aprendizagem profunda para super-resolução A propor um modelo não-sequencial, usando uma camada de *skip* conectando sua camada de entrada à sua última camada.

A arquitetura dessa rede neural é muito semelhante à da SRCNN. A interpolação é feita antes da camada de entrada da rede neural. A camada de entrada é seguida por N camadas de convolução com o *kernel* $(3, 3)$, no artigo original o autor sugere o uso de $N = 20$ camadas convolucionais, a arquitetura desta rede é mostrada na figura 3.

Figura 3 – Arquitetura original da VDSR.



Fonte:Imagem de autoria própria (2019).

Essa rede já produziu alguns resultados na área de extração de texto como mostrado por Zhang et al.[6].

3.3 *Deeply-Recursive Convolutional Network*

A *deeply-recursive convolutional network* (DRCN)[13] foi proposta por Kim et al. na CVPR 2016. Esta foi a primeira rede neural de aprendizagem profunda voltada para super resolução a usar uma camada recursiva. Também essa foi a primeira dentre as técnicas de super resolução a encarar o problema do "gradiente de fuga"(vindo do inglês *vanishing gradient problem*).[30].

Como no trabalho original[13], o DRCN é dividido em três sub-redes, como mostrado na figura 4:

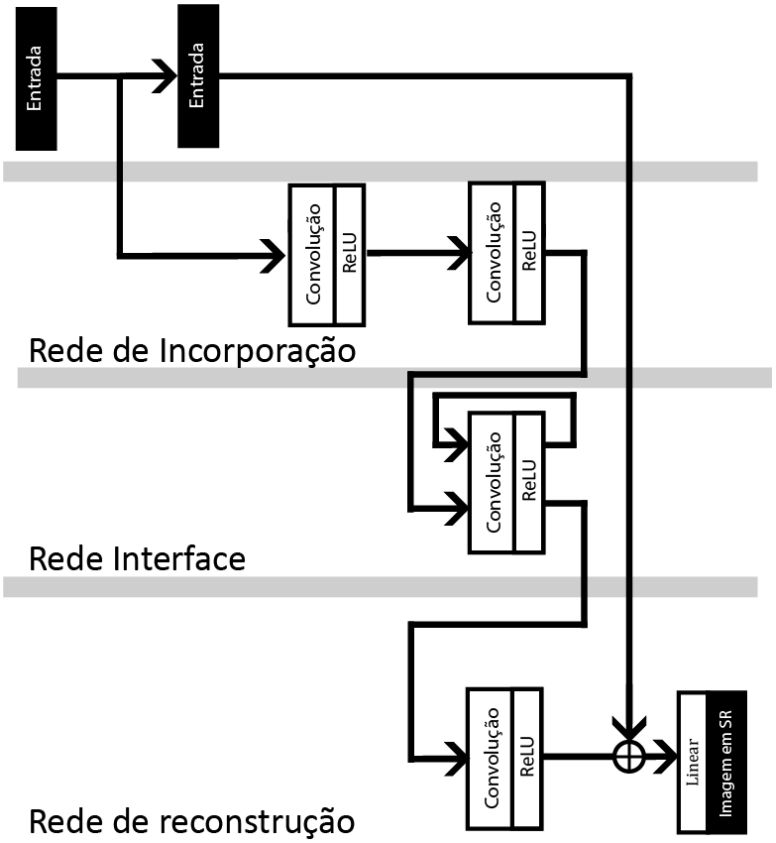
1. **Rede de Incorporação (do ingês *Embedding Net*)** Este bloco transforma a camada de entrada nos mapas de características.

2. **Rede Interface (do inglês *Inference Net*)** o bloco principal do DRCN. Ele é composto por um único bloco recursivo com um kernel maior que *(1,1)* seguido por uma ativação de ReLU. Este bloco é responsável pela análise dos mapas de características gerados na rede de incorporação.

3. **Rede de reconstrução (do inglês *Reconstruction Net*)** Este é o último bloco da DRCN, responsável pela união dos diferentes mapas de características na rede de inferência.

A arquitetura da DRCN pode ser vista na figura 4, e como representado nela existe uma conexão de *skip* entre a entrada e a saída, assim como a VDSR .

Figura 4 – Arquitetura original da DRCN.



Fonte:Imagem de autoria própria (2019).

3.4 *Fast Super Resolution Convolutional Network*

Baseada na SRCNN[7] a *Fast Super Resolution Convolutional Network* (FSRCNN) foi proposta por Dong et al.[9]. A FSRCNN foi muito importante devido a velocidade apresentada, superando modelos anteriores. Outra contribuição deste trabalho é o fato de que essa foi a primeira rede de SR à incorporar a *Parametric rectified linear unit*(PReLU)[31].

A arquitetura da FSRCNN, apresentada na figura 5, é divida em 5 blocos[9], cada uma das camadas possui é seguida de uma camada de ativação PreLU, com a exceção da última. Seus blocos são descritos como:

1. **Extração de características (do inglês, *Feature extraction*)**: Este bloco extrai as características da imagem de baixa resolução, a partir de uma convolução. Tal processo se dá de forma análoga à interpolação no SRCNN.

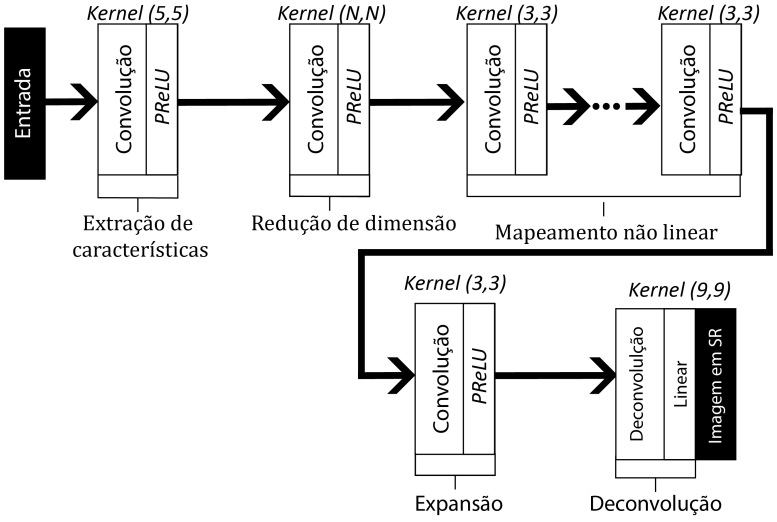
2. **Redução de dimensão (do inglês, *Shrinking*)**: Responsável pela redução da dimensão da saída da camada de extração de características na FSRCNN, já que imagens de baixa resolução produzem mapas de características grandes.

3. **Mapeamento não linear (do inglês, *Non-linear mappings*)**: Este bloco contém as principais camadas da FSRCNN. Ele é responsável pelo mapeamento da imagem de SR.

4. **Expansão (do inglês, *Expanding*)** Bloco responsável pela descriptografia dos mapas de características modificados, agindo como o processo inverso da camada de redução de dimensão, visando que a imagem de SR tenha as mesmas características da imagem de baixa resolução, porém aprimoradas pelo bloco de mapeamento não linear.

5. **Deconvolução (do inglês, *Deconvolution*)** Este é a último bloco da FSRCNN, ele é responsável pelo processo de deconvolução, assim como pode vir a ser o responsável pelo aumento nas dimensões.

Figura 5 – Arquitetura original da FSRCNN.



Fonte:Imagem de autoria própria (2019).

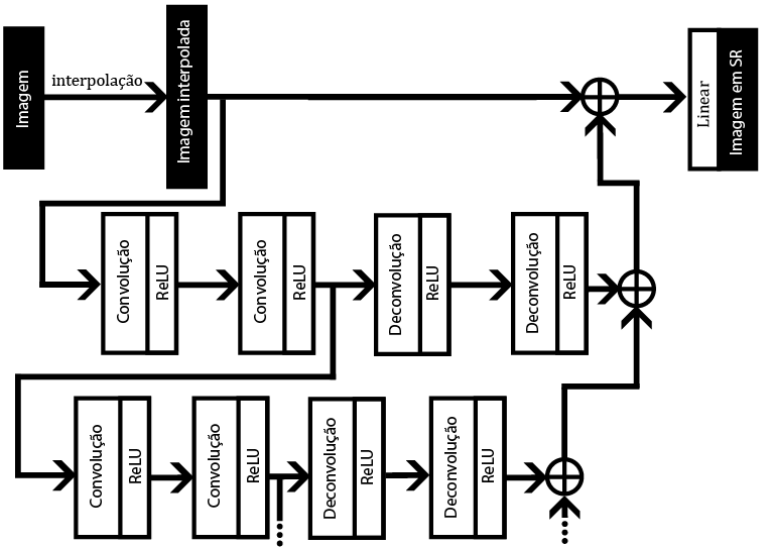
3.5 *Residual Encoder-Decoder Network*

Proposta por Mao et al.[14] a *Residual Encoder-Decoder Network* (RED-Net) é baseada na seguinte hipótese: Se uma imagem com ruído pode ser representado pela equação 3.1. (Onde N é o ruído, H é o erro da degradação, x é a versão sem ruído.), então é possível para uma rede de aprendizagem profunda, uma vez treinada para reconhecer o erro, ser capaz de removê-lo.A principal contribuição da RED-Net é o fato de que esse foi o primeiro *auto-encoder* dentre as técnicas de super resolução.

$$y = H(x) + n \tag{3.1}$$

A arquitetura dessa rede é composta por camadas de convolução ligadas a camadas de deconvolução de forma simétrica seguidas por ativações ReLU, como mostra a figura 6.

Figura 6 – Arquitetura original da RED-net.



Fonte:Imagem de autoria própria (2019).

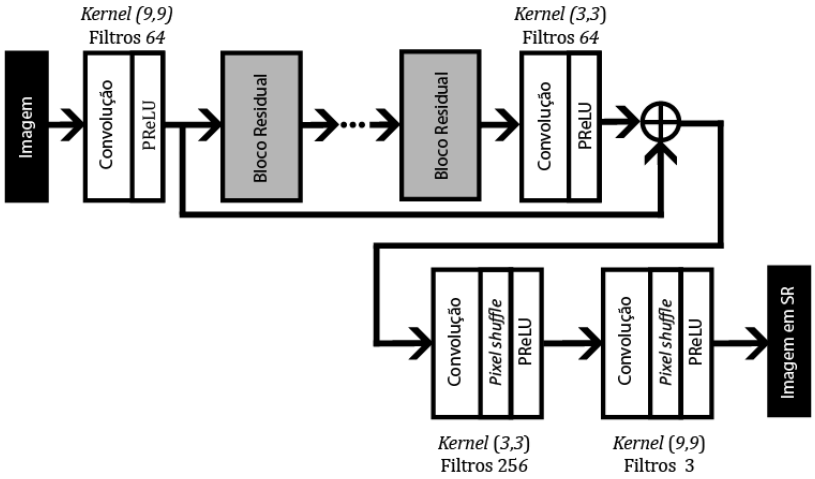
3.6 *Super Resolution Generative Adversarial Network*

Proposta no CVPR 2017 por Ledig et al. a *Super Resolution Generative Adversarial Network*(SRGAN)[15],foi a primeira GAN entre as técnicas de SR. Essa mudança foi motivada pela perda de detalhes causada pela função de perda, o erro quadrático médio (MSE)[22, 32]. Devido a essa falha, Ledig et al. propôs uma arquitetura usando de uma GAN aproveitando-se da perda perceptiva (Do inglês *Perceptual Loss*)[33].

A arquitetura da SRGAN é composta por um gerador, um discriminador e um extrator de características, que consiste em uma rede VGG. O gerador da SRGAN é uma rede residual[34], é o componente da GAN que recebe como entrada as imagens em baixa

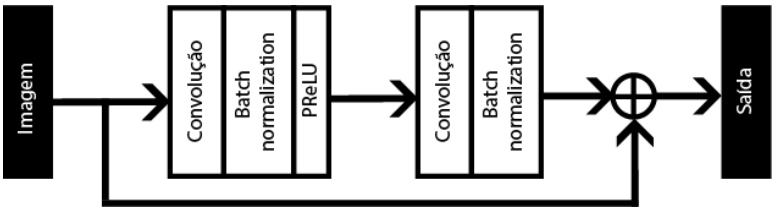
resolução usando de N blocos residuais para "modificar"a imagem de baixa resolução, e camadas de *pixel shuffle* para ampliar a imagem. O gerador dessa rede é representado na figura 7. Enquanto os blocos residuais são representados pela figura 8.

Figura 7 – Arquitetura original do gerador da SRGAN.



Fonte:Imagem de autoria própria (2019).

Figura 8 – Arquitetura original do bloco residual usado na SRGAN.

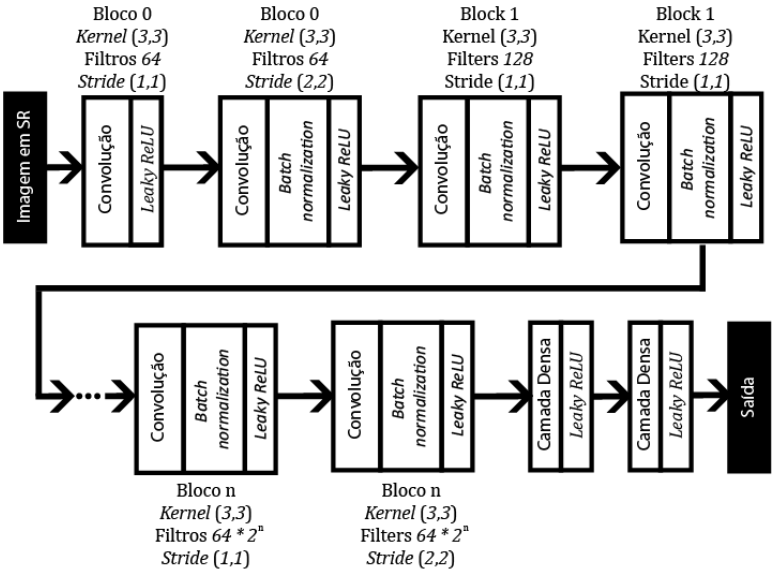


Fonte:Imagem de autoria própria (2019).

Baseado em Radford et al.[35] o discriminador da SRGAN funciona extraindo as características através de uma combinação de convolução, *batch normalization* [36] e *Leaky ReLU* [37]. Esses blocos variam ao longo da rede. Onde o primeiro bloco da rede não tem a camada de *batch normalization*, enquanto que nos blocos onde o número de filtros é duplicado é aplicado um *stride* de $(2, 2)$ para reduzir a dimensão da imagem. Após as n interações , os mapas de características gerados por este processo são usados como entrada para uma camada densa, a fim de determinar a probabilidade de esta imagem ser realista. Este discriminador pode ser visto na figura 9.

Esta nova abordagem na função de perda do SRGAN foi uma contribuição importante, a *perceptual loss* [33] provou alcançar melhores resultados do que as funções de perda de anterior. Esta perda considera as características calculadas usando um extrator

Figura 9 – Arquitetura original do discriminador da SRGAN.



Fonte:Imagem de autoria própria (2019).

de características para alcançar imagens mais realistas, no original é usada uma rede VGG[29]. Esta função perdida pode ser vista na equação. 3.2

$$l^{sr} = l^{sr}_{x} + 10^{-3} l^{sr}_{gen} \tag{3.2}$$

Como visto na equação 3.2 existem 2 componentes na *perceptual loss*: o *content loss* representado por I^{sr}_{x} e a *adversarial loss* representada por l^{sr}_{gen}. O *content loss* é calculado com base nas características extraídas pelo do VGG e o resultado do MSE, conforme mostrado em equação 3.3 e 3.4.

$$l^{SR}_{MSE} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I^{HR}_{x,y} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{3.3}$$

$$\begin{aligned} l^{SR}_{VGG_{i,j}} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{LR})_{x,y} \\ -\phi_{i,j}(G_{\theta_G}(I^{HR})_{x,y})^2 \end{aligned} \tag{3.4}$$

Na equação 3.3, r representa o fator de redimensionamento enquanto W e H representam as dimensões da imagem (largura e altura) e $I^{HR}_{x,y}$, $G_{\theta_G}(I^{LR})_{x,y}$ representam respectivamente os pixeis na imagem de alta resolução e os pixeis da imagem de super-resolução gerada. Na equação 3.4 são usados símbolos similares, mas esses recursos são calculados por cada convolução j e *max-pooling* i na VGG.

A *adversarial loss* é calculada sobre a saída do discriminador sobre todas as amostras, como mostrado na Equação 3.5, onde o $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ representa a probabilidade da imagem gerada $G_{\theta_G}(I^{LR})$ seja realista. Esta função foi escolhida sobre a original, representada por $\log|1 - D_{\theta_D}(G_{\theta_G}(I^{LR}))|$ para ter um melhor comportamento de gradiente.

$$l^{SR}_{GEN} = \sum_{N=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{3.5}$$

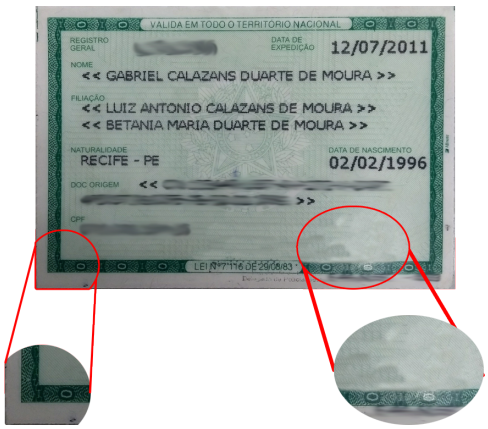
Outra contribuição da SRGAN foi a análise realizada mostrando que mesmo tendo uma taxa de PSNR menor do que outros métodos, os resultados da SRGAN conseguiu uma taxa de MOS maior do que outros métodos de super resolução

4 Experimentos

4.1 Base de dados

Durante nossas primeiras abordagens, tentamos usar um aplicativo móvel para criar esse novo conjunto de dados. no entanto, devido a ruído nessas imagens, algumas redes aprenderam como reproduzir e aplicar esse ruído na imagem SR, uma amostra desse conjunto de dados pode ser visto na figura 10. Por isso, decidimos usar um cenário digitalizado pelo escâner.

Figura 10 – Amostra da base de dados inicial e dos ruídos presentes nas imagens.



Fonte:Imagem de autoria própria (2019).

Para obtermos uma base de dados mais realista, nós utilizamos uma base privada de documentos brasileiros[1] para treinar e validar nossas redes. Esta base de dados privada contém 954 imagens com 200 DPI, mas com sistema de cores, larguras e alturas diferentes. A altura das imagens nessa base variam de 2800 a 377 pixeis, as larguras nessa base variam de 1700 a 546 pixeis e o conjunto de cores em RGB e em escala de cinza.

4.2 Pré-processamento dos dados

Para o teste e treino, selecionamos 642 (67%) imagens de treino, 145 (15%) imagens de teste e 167 (17%) imagens para validação dos resultados. Assim como haviam poucas imagens foi usado de *data augmentation*, girando a imagem em uma taxa de 15º variando de 0º a 360º e invertendo a imagem em 2 orientações (vertical and horizontal). Devido a essas transformações a quantidade de imagens foi multiplicada por 27x, outros tipos de aumento de dados foram não usado, pois seria diferente do cenário desejado.

[1] Devido a informações sigilosa presente nessa base de dados não podemos prover nenhum exemplo dela.

Para abordar imagens com tamanhos diferentes sem fazer nenhuma redução de tamanho, usamos de uma estratégia de divisão. Primeiramente foi adicionada uma borda branca para que as dimensões das imagens pudesse ser um múltiplo de 256, em seguida, as imagens foram dividas em sub imagens de 256x256 pixeis. Ao fazer o processo descrito, reduzimos o uso de memória da unidade de processo gráfico durante as etapas de treinamento e teste.

Para obter imagens de baixa resolução durante o treinamento da rede, as sub-imagens foram reduzidas de acordo com o redimensionamento desjeado da rede: tiveram seu tamanho reduzido pela metade para as redes de 2x e reduzidas a um quarto para redes de 4x. Junto a isso foi feita uma interpolação bicúbica [19] antes da entrada nas redes: SRCNN, VDSR, DRCN e RED-Net.

4.3 Detalhes da implementação das redes

Para padronizar o treino as seguintes decisões foram realizadas:

- **Otimizador**: Durante o processo de treino o otimizador Adam[38] foi utilizado com uma taxa de aprendizagem de $3e^{-3}$, $beta_1 = 0,9$ e $beta_2 = 0,9999$.

- **Função de perda**: Foi utilizado o erro quadrático médio na maioria das redes, tal equação é descrita por 4.1 nessas equações $I^{HR}_{x,y}$ representa o pixel da imagem em alta resolução(x,y), $I^{SR}_{x,y}$ representa o pixel da imagem em super resolução (x,y). Entretanto, foi usado a função de perda perceptiva[33] na SRGAN descrita no capitulo 3 na equação 3.5

$$MSE = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}(I^{HR}_{x,y} - (I^{SR})_{x,y})^2 \tag{4.1}$$

- **Métricas**: Durante o processo de treino foram usadas 2 métricas: A distância de Jaccard e o PSNR essas métricas são descritas nas equações 4.2 e 4.3, nestas equações $nbits$ representa o número de bits na imagem original, I^{SR} representa a imagem em SR enquanto I^{HR} representa a imagem em alta resolução.

  A distância de Jaccard também pode ser interpretada como a interseção dividida pela união. Esta métrica define o quão similar o conjunto de pixeis representados por I^{SR} é similar ao conjunto de pixeis representado por I^{SR}

$$dist\hat{a}nciadeJaccard = \frac{|I^{HR} \cap I^{SR}|}{|I^{HR} \cup I^{SR}|} \tag{4.2}$$

$$PSNR = 10log\frac{(2^{nBits} - 1)^2}{MSE} \tag{4.3}$$

Além do uso de um otimizador diferente em algumas redes, reduzimos a arquitetura do SRGAN e repetimos algumas camadas no FSRCNN. No gerador da SRGAN foram removidas as 3 últimas camadas e reduzimos o número de blocos de amostragem em 1 no discriminador para realizar uma ampliação de 2x. No FSRCNN foi adicionado uma camada de deconvolução no final, para a realização de ampliações de 4x.

Durante os testes, notamos que o VDSR treinado com um otimizador Adam e sem a camada *skip* têm melhores resultados que o VDSR com a camada *skip*, como mostra a seção de resultados. Por outro lado, não houve outras modificações no SRCNN, DRCN e RED-Net. Durante este trabalho, todas as redes foram desenvolvidas usando as bibliotecas em Python do Keras [39] rodando sobre o TensorFlow[40].

4.4 Treino e testes das redes

Durante o processo de treino foi usado um lote de 32 imagens, no qual cada uma dessas imagens foi dividida de acordo com o processo descrito na secção de pré processamento, criando uma média de 192 imagens por lote. Mas no caso de SRGAN esse lote foi reduzido para 5 imagens (30 sub imagens) uma vez que a unidade de processamento gráfico não conseguia suportar uma quantidade maior de imagens.

Cada época usou 1000 imagens e todas as redes foram treinadas por 350 épocas. Demorando em média 2 dias para treinar cada rede. Durante nosso treino usamos ambos o *Google cloud platform*[2] equipado com uma NVIDIA Tesla P100 e um computador de mesa com uma Geforce 1060 6gb.

Durante os testes, toda a base de testes foi utilizada. Nós utilizamos de uma abordagem com mais de uma métrica. Foram utilizadas a distância de Jaccard e a PSNR. O monitoramento das mesmas durante o treinamento foi feito conforme a seguir: Foi considerado que havia um ganho no treino apenas quando a distância de Jaccard estivesse abaixo de 2.00 e o PSNR maior que os antigos do que os valores antigos, essa métrica foi usada a fim de evitar imagens monocromáticas.

4.5 Pós Processamento

Após a etapa de treinamento e teste, a base de validação foi preparada para ser utilizada nos testes de reconhecimento óptico de caracteres. Inicialmente a base de validação foi rotulada. Após a rotulação foram geradas imagens com 200 DPI, 300 DPI e 400 DPI. Para gerar as imagens necessárias foi utilizada de uma interpolação bicúbica [19] para reduzir as imagens por fatores de 2x, 1,5x, 2x, após o que fizemos a ampliação de todas as imagens com um fator de 2x, 2x, 4x. Fazendo isso para as três imagens de resolução do

[2] https://console.cloud.google.com acessado in 23/06/2019

alvo. Depois de aumentar o escalonamento usando as redes e convertê-las para a escala de cinza, usamos uma técnica de equalização de histograma. Também foi preparada uma base interpolada, aplicando a interpolação bicúbica[19] na imagem original para ter uma saída com o mesmo tamanho da saída da rede neural.

Durante nossos testes também aplicamos a binarização de Sauvola [41], com $W = 60$ e $K = 0.01$, para avaliar o desempenho dos mecanismos de OCR no cenário de imagens binárias.

4.6 Avaliação do OCR

Durante os testes, devido aos resultados de trabalhos anteriores [5, 42], selecionamos dois produtos de OCR populares: O Tesseract OCR (versão 4.0) e o ABBYY *fine reader* (versão 12.0). Usamos a configuração padrão no Tesseract enquanto na ABBYY desativamos a opção "Corrigir resolução da imagem". A entrada dos dois mecanismos OCR foi a imagem pós-processada, tanto a imagem cinza quanto a imagem binária. Para calcular a taxa de erro de caracteres e a taxa de erro de palavras foi usada a distância de Levenshtein [43].

5 Resultados

Após os experimentos, foram medidas as taxa de erro por letras, como mostrado na tabela 1, e a taxa de erro por palavras, conforme mostrado na tabela 2.

5.1 Resultados do OCR

Ambas as tabelas mostram que a maioria das redes de SR conseguem diminuir a taxa de erro. Analisando as tabelas, podemos afirmar que:

- Quando comparando a saída gerada em escala de cinza de 200 DPI, tanto a taxa por letra quanto a taxa de erro da palavra aproximam-se da original. Além disso, algumas redes ficaram muito próximas da taxa original. Assim, é possível supor que, se a imagem for reduzida em 2x, a imagem SR gerada por uma dessas redes estará próxima da imagem original, usando como base a métrica de precisão do OCR.
- Comparando a saída gerada em escala de cinza de 300 DPI, a diminuição na taxa de erro é maior que a imagem de 200 DPI. Embora em alguns casos as imagens em 400 DPI tenham obtido resultados mais baixos, podemos supor que este é o melhor cenário devido à perda quando comparado com outras resoluções. Isso ocorre porque alguns dos mecanismos de OCR são treinados para usar uma imagem de 300 DPI. [44]
- Comparando a saída gerada em escala de cinza de 400 DPI, a taxa de erro é mais baixa que a original. Entretanto, algumas vezes as imagens de 300 DPI têm taxas menores que está, como esperado. Isso porque em algum momento durante a redução da imagem, houve perda de dados, de modo que nem as redes puderam restaurar esses detalhes. Mas mesmo com essa perda de dados, os resultados provaram estar próximos da imagem original.

5.2 Resultados nas imagens

Os resultados da PSNR e a distância de Jaccard são ilustrados na tabela 3. Esses resultados foram calculados sobre a imagem pós-processada no cenário de 200 DPI.

Como mostrado na tabela 3, a distância de Jaccard é similar em todos os métodos, tendo uma pequena variação entre eles, o que implica que as imagens são semelhantes. No entanto, os valores de PSNR têm uma certa variação entre eles, isto é provavelmente

Tabela 1 – Taxa de erro por letras em diferentes cenários

Taxa de erro de letras (desvio padrão da taxa)				
		200 DPI	300 DPI	400 DPI
Tesseract (imagem binária)	Interpolação bicúbica[19]	63,62% (0,3369)	51,15% (0,2954)	52,34% (0,2362)
	DRCN[13]	54,09% (0,3231)	48,47% (0,2946)	45,99% (0,2781)
	RED-NET[14]	52,86% (0,3267)	47,58% (0,3156)	44,75% (0,2460)
	FSRCNN[9]	53,19% (0,3335)	49,37% (0,3016)	49,02% (0,2824)
	SRCNN[7]	53,75% (0,3364)	47,22% (0,3068)	43,48% (0,2245)
	SRGAN[15]	**46,65%** (0,3024)	**43,02%** (0,3043)	48,75% (0,2911)
	VDSR modificada	58,81% (0,3345)	46,32% (0,3024)	**42,97%** (0,2553)
	VDSR usando camada de *skip* [8]	73,71% (0,2604)	60,23% (0,2636)	58,55% (0,2933)
	Imagem original (200 DPI)		50,45% (0,3417)	
	Imagem em baixa resolução (100 DPI)		68,58% (0,2974)	
ABBYY (imagem binária)	Interpolação bicúbica[19]	70,92% (0,2973)	62,23% (0,3477)	59,17% (0,3047)
	DRCN[13]	62,24% (0,3733)	61,62% (0,3649)	52,33% (0,3362)
	RED-NET[14]	60,04% (0,3746)	60,83% (0,3651)	51,84% (0,3336)
	FSRCNN[9]	64,42% (0,3626)	62,15% (0,3721)	54,23% (0,3339)
	SRCNN[7]	62,65% (0,3707)	58,75% (0,3648)	51,64% (0,3313)
	SRGAN[15]	**56,61%** (0,3494)	**50,33%** (0,3146)	59,56% (0,3606)
	VDSR modificada	64,46% (0,3481)	58,67% (0,3720)	**50,49%**(0,3377)
	VDSR usando camada de *skip* [8]	84,68% (0,2100)	70,36% (0,2854)	66,11% (0,3082)
	Imagem original (200 DPI)		58,50% (0,3756)	
	Imagem em baixa resolução (100 DPI)		75,34% (0,2878)	
Tesseract (usando imagem cinza)	Interpolação bicúbica[19]	39,01% (0,2331)	30,71% (0,2157)	35,73% (0,2064)
	DRCN[13]	33,85% (0,2281)	28,69% (0,3066)	29,32% (0,2131)
	RED-NET[14]	32,25% (0,2256)	28,41% (0,3055)	30,37% (0,2108)
	FSRCNN[9]	34,40% (0,2322)	29,11% (0,3175)	31,54% (0,2320)
	SRCNN[7]	34,01% (0,2329)	28,52% (0,3017)	30,55% (0,2096)
	SRGAN[15]	**32,11%** (0,2246)	28,36% (0,2686)	**27,76%** (0,2020)
	VDSR modificada	36,70% (0,2330)	**28,09%** (0,2836)	29,57% (0,2107)
	VDSR usando camada de *skip* [8]	70,97% (0,2668)	53,86% (0,2688)	49,90% (0,2960)
	Imagem original (200 DPI)		30,10% (0,2045)	
	Imagem em baixa resolução (100 DPI)		53,35% (0,2741)	
ABBYY (usando imagem cinza)	Interpolação bicúbica[19]	36,16% (0,2992)	33,02% (0,2403)	36,36% (0,2604)
	DRCN[13]	40,36% (0,3468)	35,35% (0,3168)	31,40% (0,2398)
	RED-NET[14]	37,89% (0,3308)	34,09% (0,3109)	35,24% (0,2821)
	FSRCNN[9]	37,10% (0,3213)	35,06% (0,3061)	33,84% (0,2570)
	SRCNN[7]	38,56% (0,3335)	33,31% (0,2916)	**29,04%** (0,2147)
	SRGAN[15]	**35,13%** (0,3201)	**31,79%** (0,2719)	30,49% (0,2528)
	VDSR modificada	42,61% (0,3392)	32,67% (0,2884)	31,96% (0,2469)
	VDSR usando camada de *skip* [8]	61,68%(0,3341)	51,72% (0,3079)	48,52% (0,3253)
	Imagem original (200 DPI)		27,82% (0,3189)	
	Imagem em baixa resolução (100 DPI)		49,23% (0,3409)	

devido à quantidade de textura no fundo das imagens, como algumas das imagens têm mais detalhes em seu plano de fundo do que outras.

Tabela 2 – Taxa de erro por letras em diferentes cenários

	Taxa de erro de palavras			
		200 DPI	300 DPI	400 DPI
Tesseract (usando imagem binária)	Interpolação bicúbica[19]	65,72% (0,3110)	55,42% (0,2612)	56,52% (0,2038)
	DRCN[13]	56,09% (0,3037)	52,54% (0,2663)	50,68% (0,2528)
	RED-NET[14]	55,46% (0,3049)	51,77% (0,2899)	49,95% (0,2250)
	FSRCNN[9]	56,11% (0,3108)	52,97% (0,2750)	53,66% (0,2556)
	SRCNN[7]	56,48% (0,3127)	51,14% (0,2765)	48,34% (0,2027)
	SRGAN[15]	**50,41%** (0,2822)	**46,72%** (0,2805)	52,73% (0,2632)
	VDSR modificada	61,29% (0,3112)	50,58% (0,2741)	**48,02%** (0,2331)
	VDSR usando a *skip layer*[8]	75,51% (0,2365)	63,66% (0,2384)	61,09% (0,2670)
	Imagem original (200 DPI)		53,47% (0,2950)	
	Imagem em baixa resolução (100 DPI)		71,23% (0,3147)	
ABBYY (usando imagem binária)	Interpolação bicúbica[19]	71,87% (0,2776)	65,41% (0,3091)	63,88% (0,2656)
	DRCN[13]	64,82% (0,3351)	64,50% (0,3225)	57,25% (0,2913)
	RED-NET[14]	63,21% (0,3339)	64,55% (0,3208)	57,00% (0,2942)
	FSRCNN[9]	66,80% (0,3274)	65,58% (0,3283)	58,66% (0,2931)
	SRCNN[7]	65,29% (0,3283)	62,71% (0,3214)	56,90% (0,2893)
	SRGAN[15]	**60,32%** (0,3177)	**54,21%** (0,2780)	62,78% (0,3202)
	VDSR modificada	67,00% (0,3127)	61,96% (0,3278)	**55,64**% (0,2987)
	VDSR modificada usando camada de *skip*[8]	84,00% (0,2001)	72,40%(0,2494)	68,42%(0,2723)
	Imagem original (200 DPI)		62,12% (0,3353)	
	Imagem em baixa resolução (100 DPI)		76,40% (0,2650)	
Tesseract (usando imagem cinza)	Interpolação bicúbica[19]	39,97% (0,2230)	33,28% (0,2020)	35,73% (0,1902)
	DRCN[13]	34,65% (0,2123)	31,10% (0,2803)	32,91% (0,2059)
	RED-NET[14]	**33,98%** (0,2161)	31,17% (0,2786)	33,29% (0,2023)
	FSRCNN[9]	35,89% (0,2207)	32,03% (0,2903)	35,04% (0,2185)
	SRCNN[7]	35,47% (0,2225)	31,37% (0,2780)	34,27% (0,2058)
	SRGAN[15]	34,27% (0,2167)	**30,48%** (0,2501)	**31,16%** (0,1980)
	VDSR modificada	37,64% (0,2228)	30,84% (0,2572)	33,35% (0,2053)
	VDSR usando camada de *skip*[8]	71,52% (0,2599)	56,20% (0,2402)	51,63% (0,2897)
	Imagem original (200 DPI)		31,75% (0,1860)	
	Imagem em baixa resolução (100 DPI)		54,40% (0,2563)	
ABBYY (usando imagem cinza)	Interpolação bicúbica[19]	40,69% (0,2709)	38,45% (0,2493)	42,74% (0,2382)
	DRCN[13]	44,07% (0,3145)	39,78% (0,2922)	36,42% (0,2172)
	RED-NET[14]	41,67% (0,3030)	38,62% (0,2836)	40,37% (0,2568)
	FSRCNN[9]	41,02% (0,2961)	39,33% (0,2769)	39,18% (0,2350)
	SRCNN[7]	42,84% (0,3058)	37,29% (0,2664)	**34,60%** (0,1924)
	SRGAN[15]	**39,18%** (0,2967)	**35,52%** (0,2481)	35,09% (0,2297)
	VDSR modificada	47,01% (0,3030)	37,15% (0,2623)	36,42% (0,2141)
	VDSR usando camada de *skip* [8]	63,95% (0,2999)	54,71% (0,2756)	51,30% (0,2932)
	Imagem original (200 DPI)		32,66% (0,2950)	
	Imagem em baixa resolução (100 DPI)		52,45% (0,3147)	

Tabela 3 – Os resultados das métricas das imagens

Método	PSNR Média (Desvio padrão)	Distancia de Jaccard Média (Desvio padrão)
Interpolação bicúbica[19]	18,90 (10,67)	1,010976 (7,60x10^{-6})
DRCN[13]	21,04 (13,39)	1,010958 (7,27x10^{-6})
RED-Net[14]	20,94 (14,01)	1,010949 (7,28x10^{-6})
FSRCNN[9]	21,14 (15,01)	1,010972 (7,29x10^{-6})
SRCNN[7]	21,04 (14,09)	1,010975 (7,43x10^{-6})
SRGAN[15]	20,33 (15,09)	1,010947 (7,24 x10^{-6})
VDSR modificada	20,40 (13,99)	1,011012 (7,44x10^{-6})
VDSR usando camada de *skip* [8]	19,38(5,17)	1,011205(8,03e^{-6})

6 Conclusão e trabalhos futuros

Neste trabalho, exploramos a eficácia de redes profundas de imagens SR em um cenário com documentos com plano de fundo contendo uma alta quantidade de textura. Primeiramente foi investigado e avaliado seis redes profundas existentes em situações cotidianas comuns. Com objetivo de avaliar a eficácia de cada modelo e entender a melhor maneira de explorar os benefícios desses modelos profundos para a tarefa de OCR. Também está comprovado que o PSNR não está fortemente relacionado à precisão do OCR. Extensas experiências em SR com modelos profundos demonstram resultados promissores para a precisão do OCR. Neste trabalho, também provamos que, em diferentes aplicações OCR, a taxa de erro tende a diminuir quando a resolução da imagem é aumentada em alguns cenários até certo ponto.

Em trabalhos futuros, avaliaremos o desempenho de redes recentes como a rede de super resolução progreciva (ProSR, do inglês *Progressive super resolution*) [45], a rede de super resolução com retroalimentação (SRFBN, do inglês *Super resolution feedback network*) [46], a SRGAN melhorada (ESRGAN, do inglês *Enhanced SRGAN*) [47] e a rede residual orientadas a canais (RCAN [48], do inglês *Residual Channel Attention Networks*). Também há uma alta probabilidade de desenvolvermos uma nova arquitetura de aprendizagem profunda que seja capaz de lidar com os desafios das imagens SR para aplicações OCR.

Referências

[1] KASPARAOV, G. *Don't fear intelligent machines. Work with them.* 2017. Acessado em 23/06/2019. Disponível em: <https://www.youtube.com/watch?v=NP8xt8o4_5Q&t30s/>. Citado na página 5.

[2] UMEHARA, K.; OTA, J.; ISHIDA, T. Application of super-resolution convolutional neural network for enhancing image resolution in chest ct. *Journal of Digital Imaging*, v. 31, 2018. Citado 2 vezes nas páginas 13 e 18.

[3] MAGADOMI. *Waifu2x.* 2015. Acessado em 06/06/2019. Disponível em: <http://waifu2x.udp.jp>. Citado 3 vezes nas páginas 13, 14 e 18.

[4] PANDEY, R.; MAIYA, S.; RAMAKRISHNAN, A. A new approach for upscaling document images for improving their quality. In: *2017 14th IEEE India Council International Conference (INDICON).* [S.l.: s.n.], 2017. Citado 4 vezes nas páginas 13, 14, 17 e 18.

[5] LAT, A.; JAWAHAR, C. Enhancing ocr accuracy with super resolution. In: IEEE. *2018 24th International Conference on Pattern Recognition (ICPR).* [S.l.], 2018. Citado 5 vezes nas páginas 13, 14, 17, 18 e 29.

[6] Zhang, H.; Liu, D.; Xiong, Z. Cnn-based text image super-resolution tailored for ocr. In: *2017 IEEE Visual Communications and Image Processing (VCIP).* [S.l.: s.n.], 2017. Citado 5 vezes nas páginas 13, 14, 17, 18 e 19.

[7] DONG, C. et al. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 38, n. 2, 2015. Citado 9 vezes nas páginas 13, 14, 16, 18, 19, 20, 31, 32 e 33.

[8] KIM, J.; LEE, J. K.; LEE, K. M. Accurate image super-resolution using very deep convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* [S.l.: s.n.], 2016. Citado 8 vezes nas páginas 13, 14, 16, 17, 19, 31, 32 e 33.

[9] DONG, C.; LOY, C. C.; TANG, X. Accelerating the super-resolution convolutional neural network. In: SPRINGER. *European conference on computer vision.* [S.l.], 2016. Citado 8 vezes nas páginas 13, 14, 16, 20, 21, 31, 32 e 33.

[10] ASAD, F. et al. High performance ocr for camera-captured blurred documents with lstm networks. In: IEEE. *2016 12th IAPR Workshop on Document Analysis Systems (DAS).* [S.l.], 2016. Citado na página 13.

[11] CHINARA, C. et al. A novel approach to skew-detection and correction of english alphabets for ocr. In: IEEE. *2012 IEEE Student Conference on Research and Development (SCOReD)*. [S.l.], 2012. Citado na página 13.

[12] Mande Shen; Hansheng Lei. Improving ocr performance with background image elimination. In: *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. [S.l.: s.n.], 2015. Citado na página 13.

[13] KIM, J.; LEE, J. K.; LEE, K. M. Deeply-recursive convolutional network for image super-resolution. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. Citado 7 vezes nas páginas 13, 14, 16, 19, 31, 32 e 33.

[14] MAO, X.-J.; SHEN, C.; YANG, Y.-B. Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921*, 2016. Citado 7 vezes nas páginas 13, 14, 16, 22, 31, 32 e 33.

[15] LEDIG, C. et al. Photo-realistic single image super-resolution using a generative adversarial network. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. Citado 7 vezes nas páginas 13, 14, 16, 22, 31, 32 e 33.

[16] SUSAN, S.; DEVI, K. R. Text area segmentation from document images by novel adaptive thresholding and template matching using texture cues. *Pattern Analysis and Applications*, Springer, p. 1–13, 2019. Citado na página 13.

[17] DUCHON, C. E. Lanczos filtering in one and two dimensions. *Journal of applied meteorology*, v. 18, n. 8, 1979. Citado na página 16.

[18] LI, X.; ORCHARD, M. T. New edge-directed interpolation. *IEEE transactions on image processing*, IEEE, v. 10, n. 10, 2001. Citado na página 16.

[19] KEYS, R. G. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust., Speech, Signal Process*, p. 1153–1160, 1981. Citado 7 vezes nas páginas 16, 27, 28, 29, 31, 32 e 33.

[20] KAWULOK, M. et al. Deep learning for multiple-image super-resolution. *arXiv preprint arXiv:1903.00440*, 2019. Citado na página 16.

[21] FARSIU, S. et al. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, IEEE, v. 13, n. 10, 2004. Citado na página 16.

[22] MATHIEU, M.; COUPRIE, C.; LECUN, Y. *Deep multi-scale video prediction beyond mean square error*. 2015. Citado 2 vezes nas páginas 16 e 22.

[23] FREEMAN, W. T.; JONES, T. R.; PASZTOR, E. C. Example-based super-resolution. *IEEE Computer graphics and Applications*, IEEE, n. 2, 2002. Citado na página 16.

[24] GLASNER, D.; BAGON, S.; IRANI, M. Super-resolution from a single image. In: *Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan*. [S.l.: s.n.], 2009. Citado na página 16.

[25] IRANI, M.; PELEG, S. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, Elsevier, v. 53, n. 3, 1991. Citado na página 16.

[26] LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Taipei, Taiwan, v. 86, n. 11, 1998. Citado na página 16.

[27] GOODFELLOW, I. et al. Generative adversarial nets. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2014. Citado na página 16.

[28] ANWAR, S.; KHAN, S.; BARNES, N. A deep journey into super-resolution: A survey. *arXiv preprint arXiv:1904.07523*, 2019. Citado na página 18.

[29] SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Citado 2 vezes nas páginas 19 e 24.

[30] BENGIO, Y. et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, v. 5, n. 2, 1994. Citado na página 19.

[31] HE, K. et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. Citado na página 20.

[32] WANG, Z.; BOVIK, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, IEEE, v. 26, n. 1, 2009. Citado na página 22.

[33] JOHNSON, J.; ALAHI, A.; FEI-FEI, L. Perceptual losses for real-time style transfer and super-resolution. In: SPRINGER. *European conference on computer vision*. [S.l.], 2016. Citado 3 vezes nas páginas 22, 23 e 27.

[34] HE, K. et al. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016. Disponível em: <http://dx.doi.org/10.1109/CVPR.2016.90>. Citado na página 22.

[35] RADFORD, A.; METZ, L.; CHINTALA, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. Citado na página 23.

[36] IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. Citado na página 23.

[37] XU, B. et al. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. Citado na página 23.

[38] KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Citado na página 27.

[39] CHOLLET, F. et al. *Keras*. 2015. `https://keras.io`. Acessado em 22/06/19. Citado na página 28.

[40] ABADI, M. et al. Tensorflow: A system for large-scale machine learning. In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. [S.l.: s.n.], 2016. Citado na página 28.

[41] SAUVOLA, J.; PIETIKÄINEN, M. Adaptive document image binarization. *Pattern recognition*, Elsevier, v. 33, n. 2, 2000. Citado na página 29.

[42] DONG, C. et al. Boosting optical character recognition: A super-resolution approach. *arXiv preprint arXiv:1506.02211*, 2015. Citado na página 29.

[43] LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet physics doklady*. [S.l.: s.n.], 1966. v. 10, n. 8. Citado na página 29.

[44] Google inc. *Tesseract OCR*. 2019. Https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality. Acessado em 22/06/19. Citado na página 30.

[45] YIFAN, W. et al. A fully progressive approach to single-image super-resolution. In: *CVPR Workshops*. [S.l.: s.n.], 2018. Citado na página 34.

[46] LI, Z. et al. Feedback network for image super-resolution. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. Citado na página 34.

[47] WANG, X. et al. Esrgan: Enhanced super-resolution generative adversarial networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. Citado na página 34.

[48] ZHANG, Y. et al. Image super-resolution using very deep residual channel attention networks. In: *ECCV*. [S.l.: s.n.], 2018. Citado na página 34.

Single Document Image Super Resolution Using Deep Learning for Improved OCR Performance

Gabriel Calazans Duarte de Moura
University of Pernambuco
Email: gcdm@ecomp.poli.br

Byron Leite Dantas Bezerra
University of Pernambuco
Email: byronleite@ecomp.poli.br

Estanislau Baptista Lima
University of Pernambuco
Email: ebl2@ecomp.poli.br

***Abstract*—Document falsification has been quite common in nowadays days. Those illegal practices has been opposed with computing vision applications such as optical character recognition (OCR) techniques. However, OCR's has some issues, mainly with deformation and noise, such as brightness, skew, low resolution and the image background. Those errors, which are quite common, presents a high OCR accuracy loss. Those applications becomes even more important in the digital age, where many applications have complex processes thus require a high transfer rate. In those applications there is no way to guarantee the quality of these documents, and in some cases those applications tend to deform the information making harder to OCR algorithms to archive higher rates. However, many techniques have been developed to approach some of those deformations. Image super resolution (SR) is one of these techniques that has been proposed. SR aims to transcribes a low resolution image in a fake high resolution image, and have shown a huge improvement in the last years with the advance of deep learning methods. However, we do not know the gain of these techniques when dealing with OCR techniques in a scenario where the image is not ideal for transcription due to other factors such as the texture of background the image. Therefore, in this work we propose to evaluate the use of super resolution networks in a context of documents with a high texture background to increase the OCR rate. Experimental results show that by using super resolution networks, we were able to present a decrease in the character error rate (CER) and the word error rate (WER) of two popular OCR techniques, by reaching up to 27.76% in the CER error rate, while the original image would only reach 30.10%.**

I. Introduction

Image super-resolution (SR) problem, particularly single image super-resolution, has gained increasing research attention for decades. SR techniques aims to reconstruct a high-resolution (HR) image from a single low-resolution (LR) image. Image SR is used in various computer vision applications, medicine [1], entertainment [2], including text extraction and recognition [3]–[5]. However, image SR is an ill-posed problem, since there exists multiple HR solutions for any LR input. Therefore, many studies assume that LR is a bicubic down-sampled version of HR, but other degrading factors such as blur, skew, or noise can also be considered for practical applications. To approach those problems, numerous learning based methods have been proposed to learn mappings between LR and HR image pairs.

Recently, deep learning (DL) based methods have achieved significant improvements over conventional SR methods. Among them, SRCNN [6], VDSR [7], FSRCNN [8]. Such applications have shown some achievements, due to the usage of SR images as shown by Lat et al. [4], overcoming some problems presents in LR images, which have a fewer amount of information. Therefore, is common to have HR images compressed into LR images, in order to archive higher network traffic rates or even to reduce the size of the images on disk. Due to this compression some applications, such as the optical character recognition (OCR) engines, have a decrease in their accuracy rate.

OCR engines are used to extract and manipulate image's text information. However, OCR are extremely sensitive to meta information, such as blur [9], resolution [3]–[5], skew [10] and the background [11]. In order to approach those problems many algorithms have been studied during the last years [3]–[6], [9], [10], the usage of images SR have been proposed to approach those problems [3]–[5] . However, images SR are approached in a generic way, trying to improve the detail quality of the images, Therefore, the results of a SR technique in images with a high complexity background wasn't tested in a scenario focused in the OCR. Such background is not uncommon in documents as shown in figure 1. This background have a huge impact in OCR engines, due to binarization algorithms, which are a import part in the OCR engines, which is still a challenge even in nowadays [12]

In this paper, we propose to evaluate the use of images SR in order to increase the accuracy of the OCR engine. Therefore, it was evaluated six different networks architectures: SRCNN [6], VDSR [7], FSRCNN [8], DRCN [13], RED-Net [14] and SRGAN [15], for the OCR application. The goal is to evaluate the effectiveness of those existing deep architectures, to perform OCR task. Those networks were selected due to their prior results and applications. Also it was considered the computational cost of those networks, due to physical limitations of applications that use super resolution.

During the network training was noticed some problems leading us to use a private data set of 200 dots per inch (DPI) images. That set is formed with 3 different types of Brazilian documents: driving licence(*CNH*), identity cards (*RG*) and *CPF*. Those 3 documents are the Brazilian most commons documents.

In our tests we evaluated the documents images generated by the networks in 200, 300 and 400 DPI using 2 of the most

popular OCR sofwares, Tesseract OCR[1] and the ABBYY fine reader[2]. This evaluation demonstrated that with a is possible to decrease the OCR error rate measured by the character error rate and word error rate up to 25,59%/23,24% when compared with the low resolution image in the Tesseract scenario and a 20,19%/17,85% in the ABBYY scenario. We archived a increase when compared with the high resolution image.

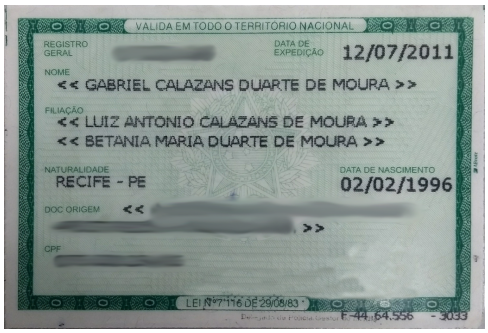


Fig. 1. A Brazilian document sample.

This work is organized as follow: in the section II is discussed the related works, briefly the super resolution focused in documents, explaining its history and also reviewing some works in this area. In section III is presented the evaluated models, explaining their architectures and showing their contributions to the SR techniques. In section IV the experiments and the methodology applied in during this work will be explained, also the implementation details will be presented. And in section V the results will be discussed.

II. Related Work

The super resolution have been approached since it conception in the 1980's. Techniques to upscale the image have been addressed in many ways [16], [17], a popular way to upscale the images images was the bicubic interpolation [18]. This interpolation has some advantages, such as a low computing cost. On the other hand the output image have normally a blurry aspect. However, many other SR methods have proven to be superior to prior techniques. Because of that, the super resolution can be grouped in two categories: the multi-image super resolution [19]–[21] and the single-image super resolution [22]–[24]. Multi-image super resolution uses the video frames or multiple photos of the same image in order to reconstruct the pixels of an image, generating a super resolution image. While the single-image super resolution uses meta information, such as edges [17] and approaching the frequency domain [16], in order to generate the SR image.

Both scenarios have presented interesting results. However, multi-image super resolution is not viable in this scenario. Because that most of the digitalized documents have just a single image, and most of them do not exist anymore or is not possible to ask the user to take more than one photo.

The single-image super resolution have show great results, especially after the work of Dong et al. [6] introducing the convolutional networks [25] in the SR problems, surpassing pior techniques.

Those networks have proved to improve the peak signal-to-noise ratio (PSNR) [6]–[8], [13], [14], overcoming other super resolution techniques. However, as shown in previous works [15], even though the PSNR rates have improved, this metric do not represent the increase in the visual quality of the super resolution image, as shown though the mean opinion score (MOS). In 2017 was proposed the usage of Generative Adversarial networks (GAN) [26] by Leding et al. [15] debuting another big step in the SR neural networks.

The impact brought by the usage of deep networks have motivated many studies. Some researches has been developed, been one of them a approach using those deep networks in order to increase the accuracy of the optical character recognition (OCR), some of those works are:

- In Pamdey et al. [3] is used a deep neural network to extract the best features in the traditional interpolation techniques such as the bicubic interpolation, the nearest neighbor and the bilinear interpolation. The combination of this features was what they called nonlinear fusion of multiple interpolations. This work has achieved some interesting results in the Tamil binary document base, increasing the accuracy by 17.89% in the best scenario.
- In Lat et al. [4] a super resolution generative adversarial network is used in the Lab color space, the training is done using 100 LR and HR pair of images captured in a controlled conditions scenarios. This scenario was evaluated using Tesseract 3.02 as well as the ABBYY fine reader, having a 21% increase of the accuracy rate OCR.
- In Zhang et al. [5] the authors used a Very deep super resolution network(VDSR) [7], During their work they proposed some modifications in the padding function and in the loss function. In this work the authors used the ICDAR 2015 TextSR data set. They reached a 78.10% OCR accuracy while the high resolution image had a 78.80% OCR accuracy

Even though those works have presented some improvement, the images used in theirs works mostly were neither a binary image or a simple background one (images there the amount of texture in the background is low). Which differs from our scenario since our images have a high amount of texture in their background.

III. Evaluated Models

In this paper six different networks were evaluated, in order to approach the document's resolution problem. We selected those networks based on their performance, computing cost and the contributions in previous works [3]–[5]. Even though there is many others super resolution networks as shown by Anwar et al. [27]

[1]Open source Tesseract project - https://github.com/tesseract-ocr/tesseract accessed in 23/06/2019

[2]ABBYY is a commercial OCR engine. https://www.abbyy.com/en-eu/ accessed in 23/06/2019

A. Super resolution convolutional neural network

Proposed by Dong et al. [6], the super resolution convolutional neural network (SRCNN) was the first deep neural network to approach the super resolution problem. Its results proved to be superior to prior super resolutions techniques as shown in the original paper [6].

The SRCNN itself is composed by three operators: a patch extraction and representation, a non-linear mapping and a reconstruction layer, as shown in figure 2. Those components may be resumed as a convolution layer followed by a ReLU activation, in such manner that the SRCNN itself do not do any upscale, thus the network focus on the noise reduction of the up-scaled image.

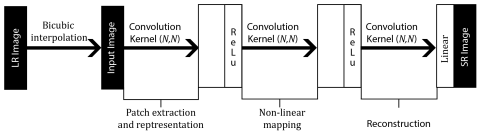


Fig. 2. The SRCNN's original architecture.

Some applications have been made using adaptations of this neural network, such as medical applications [1], and also in the entertainment area, such as the Waifu2X [2] which is used to upscale Japanese animations characters.

B. Very deep neural networks for super resolution

In 2016, Kim et al. proposed the Very deep neural networks for super resolution (VDSR) [7], based on the work of Visual Geometry Group [28] and the work of Dong et al. [6]. This work was the first super resolution deep network method to propose a non-sequential model, using a skip layer connecting its input layer to the last layer.

The architecture of this neural network is very similar to the SRCNN's. The interpolation is done before the input in the neural network. The input layer is followed by N convolution layers with (*3*,*3*) kernel, its suggest by the authors the use of $N = 20$ convolutions layers, the architecture of this network is shown in figure 3.

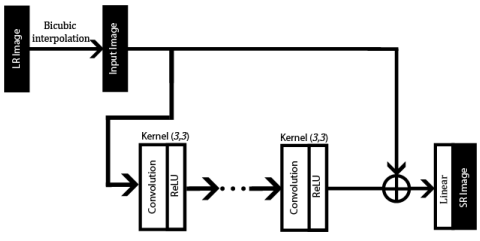


Fig. 3. The VDSR's original architecture.

This network produced some results in the text extraction area like the showed by [5].

C. Deeply-Recursive Convolutional Network

The deeply-recursive convolutional network (DRCN) was proposed by Kim et al. [13] at the 2016 CVPR. This is the first super resolution deep neural network to that used a recursive layer, thus was the first in among the SR techniques to face the vanishing gradient problem [29].

As in paper the DRCN is divided in three sub networks, as shown in figure 4:

a) Embedding Net: This block represents the input image as a set of features maps.

b) Inference Net: The main block of the DRCN, this block is composed by a single recursive block with kernel larger than (*1,1*) followed by a ReLU activation. This block is responsible by the analysis of the feature maps generated in the Embedding Net.

c) Reconstruction Net: This the last block of the DRCN, it is responsible by the union of the different feature maps in the Inference Net.

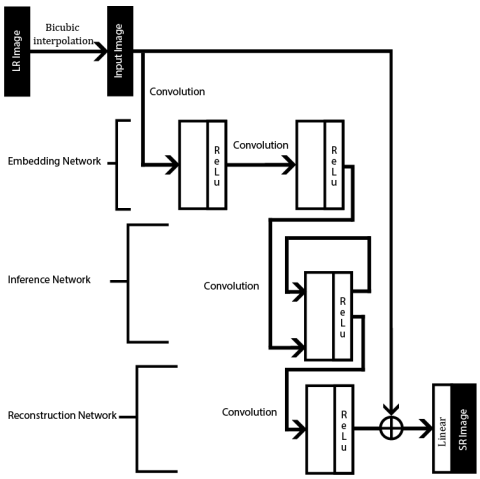


Fig. 4. The DRCN's original architecture.

D. Fast Super Resolution convolutional neural network

Based on the SRCNN [6] the Fast SRCNN (FSRCNN) was proposed by Dong et al. [8]. The FSRCNN was a important due the speed presented, surpassing prior models.Other contribution of this work is the fact that this was the first super resolution network to incorporate the PReLU [30].

The FSRCNN is divided in 5 blocks [8], each block is followed by a PReLU activation, except the last one:

a) Feature extraction: This block performs the features extraction in the low resolution image by doing a convolution,which is similar to the interpolation in the SRCNN.

b) Shrinking: This block is responsible by the dimension reduction of the Feature extraction block output in the FSRCNN, since low resolution images produce vast dimensional feature maps.

c) Non-linear mappings: This the main blocks in the FSRCNN. this block is responsible by the mapping of the SR image.

d) Expanding: This block is responsible by the decode of the modified feature maps, acting as an inverse process of the shrinking block, so the SR image will have the same features as the low resolution one but improved by the non-linear process.

e) Deconvolution: This the last block of the FSRCNN, it is responsible by the deconvolutional process. The process responsible by the up-sample of the image.

This architecture may be resumed in figure 5.

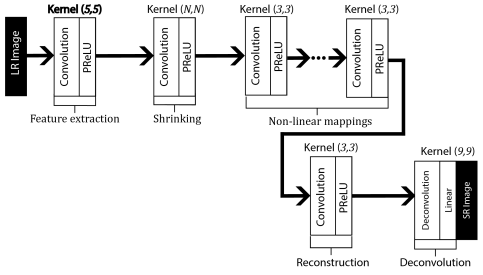


Fig. 5. The FSRCNN's original architecture.

E. Residual Encoder-Decoder Network

Proposed by Mao et al. [14] the Residual Encoder-Decoder Network (RED-Net) is based on the flowing hypotheses: If an noised image can be represented by equation1. (Where N is the noise, H is the degradation error, x is the clean version it self) Then may be possible to a deep network to remove the noise once trained to know that noise. The main contributions of the RED-Net is the fact that it was first super resolution auto encoder.

$$y = H(x) + n \tag{1}$$

The architecture of this network is composed by convolutions linked to deconvolutions in a symmetric way, followed by ReLU activation functions, as shown in the figure 6.

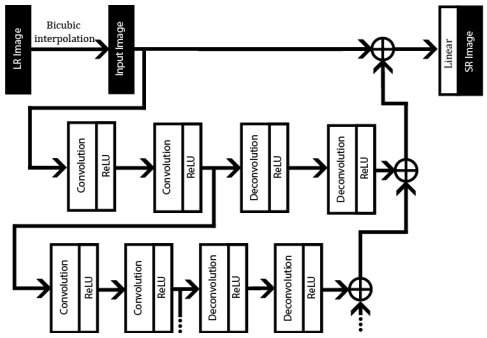


Fig. 6. The RED-Net's original architecture.

F. Super Resolution Generative Adversarial Network

Proposed in the CVPR 2017 by Ledig et al. [15], the super resolution generative adversarial network(SRGAN) was the first generative adversarial network among the SR techniques. It was motivated by the details lost caused by the loss function, the Mean Square Error(MSE) [21], [31]. Due to this failure Ledig et al. proposed a GAN architecture using the Perceptual Loss [32].

The SRGAN's architecture is composed by a generator and a discriminator. The generator of the SRGAN is a residual network [33], this part of the GAN receives as input the low resolution image as input then using N residual blocks to 'modify' the lr image image and using pixel shuffle layers to upscale them, the generator of this network may be seen in figure.7 while the Residual block of this network is shown in figure.8.

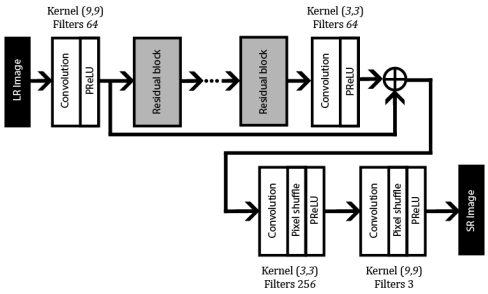


Fig. 7. The SRGAN's generator original architecture.

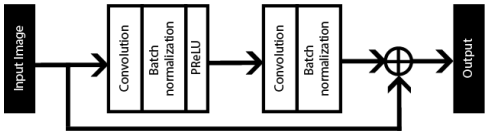


Fig. 8. The residual block used in the SRGAN.

Based on Radford et al.'s work [34] the discriminator works extracting the features through a convolution, batch normalization [35] and Leaky ReLU [36] combination. Those blocks vary over the network, the first block of the network do not have the batch normalization, while in blocks were the number of filters is doubled have a (2,2) stride in order to reduce the dimension of the image. After n interactions the feature maps generated by this process are used as input to a dense layer in order to determinate the probability of this image been realistic, this discriminator may be seen in figure 9.

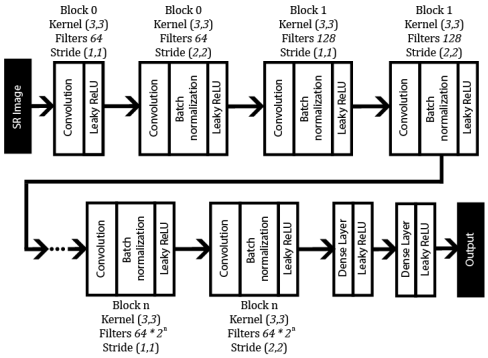


Fig. 9. The Discriminator used in the SRGAN.

This new approach in the SRGAN's loss function was a major contribution, the Perceptual Loss [32] proven to achieve

better results then prior loss functions. This loss consider the features calculated using a feature extractor in order to achieve more realistic images, in the SRGAN paper is used a VGG network [28]. This lost function may be seen in equation. 2

$$l^{sr} = l_x^{sr} + 10^3 l_{gen}^{sr} \tag{2}$$

As seen in the equation 2 there are 2 components in the lost function, the content loss represented by l_x^{sr} and the adversarial loss represented by l_{gen}^{sr}.

The content loss is calculated based on both the MSE and the VGG content as shown in equation 3 and 4.

$$l_{MSE}^{SR} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{3}$$

$$l_{VGG/i,j}^{SR} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{LR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{HR})_{x,y})^2 \tag{4}$$

In equation 3 *r* represents the upscale factor while *W* and *H* represents the dimensions of the image (width and height) and $I_{x,y}^{HR}$, $G_{\theta_G}(I^{LR})_{x,y}$ represents respectively the pixels in the high resolution image and the generated super resolution image. In equation 4 is used similar symbols but those features are calculated by each *j*-th convolution and *i*-th max-pooling.

The adversarial loss is calculated over the output of the discriminator over all samples as shown in Equation 5, where the $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ represents the probability of the generated image $G_{\theta_G}(I^{LR})$ be a realistic image. This function was chosen over the original one, represented by $\log|1 - D_{\theta_D}(G_{\theta_G}(I^{LR}))|$ in order to have a better gradient behavior.

$$l_{GEN}^{SR} = \sum_{N=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{5}$$

Another contribution of the SRGAN are their results, even having a smaller PSNR rate than other methods, the SRGAN archived an mean opinion score (MOS) higher than other super resolution methods [15].

IV. Experiments

1) Datasets: During our first approaches we tried to use a mobile application to make this new data set. However, due to some noise in those images, some networks learned how to reproduce and apply that noise in the SR image, a sample of this data set can be seen in figure 10. Therefore we decided to use a scenario digitalized by scanner.

In order to obtain a realistic dataset, we used a private Brazilian's documents[3] dataset to train and test our networks. This private dataset contains 954 images in 200 DPI but with different color spaces, width and height. The height of those images are in range of 2800 to 377 pixels, the width is in range of 1700 to 546 pixels, and the color spaces present in this data set are gray scale and rgb. In order to obtain the OCR accuracy, this dataset have been partially indexed.

[3]Due to sensitive information presented in this data set, we can not provide any sample.

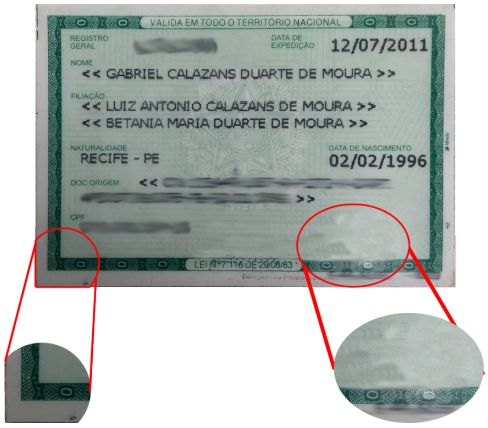


Fig. 10. An sample of this data set and some samples of the noise on the image.

2) Data pre-processing: During the test and train split we selected 642 (67%) images to train, 145 (15%) images to test and 167 (17%) images to validate the results. As were a few amount of images, we used data augmentation, rotating the image in a 15 rate in a 0 to 360 rate and flipping the image in the 2 orientations (vertical and horizontal), due to those transformations the amount of images were multiplied by 27x, other types of data augmentation were not used since it would differ from the original scenario.

In order to approach images with different sizes without doing any downscale we used a split strategy. Firstly was added a white border so the images dimensions may be a multiple of 256, then splited the images in 256x256 pixels sub-image blocks. Doing the process described we reduced the graphics process unit's memory usage during the train and test steps.

In order to obtain low resolution images during the network training, the sub-images were downscaled according to the target resize: they were downscaled by a 2x factor in order to obtain a 2x low resolution image and downscaled by a 4x factor in order to obtain a 4x low resolution image. After those downsizing was done a bicubic interpolation before the input in the SRCNN, VDSR, DRCN and RED-Net.

3) Network Implementation Details: In order to standardizing the training the following decisions have been made in the training:

- Optimizer: During the training process was used the Adam optimizer [37] with learn rate of $3e^{-3}$, $beta_1 = 0.9$ and $beta_2 = 0.9999$.
- Loss function: it was used the mean square error in the most of the networks which is described in equation.6 (In this equations $I_{x,y}^{HR}$ represents the HR image's pixel(x,y), $I_{x,y}^{SR}$ represents the SR image's pixel(x,y)), Although was used in the SRGAN the Perceptual Loss [32] described in the Evaluated models' equation.5.

$$MSE = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}(I_{x,y}^{HR} - (I^{SR})_{x,y})^2 \quad (6)$$

- Metrics: During the training process were used 2 metrics: The Jaccard distance and the peak signal-to-noise ratio (PSNR), those metrics are described in equation.7 and equation.8. In those equations nbits represents the number of bits in the original image, I^{SR} represents the SR image while I^{HR} represents the low resolution images.
  The Jaccard distance may also be interpolated as the intersection of a set divided by its union. This metric defines how much similar the set of pixels I^{SR} is similar to the set I^{SR}

$$Jaccarddistance = \frac{|I^{HR} \cap I^{SR}|}{|I^{HR} \cup I^{SR}|} \quad (7)$$

$$PSNR = 10log\frac{(2^{nBits} - 1)^2}{MSE} \quad (8)$$

Besides the usage of a different optimizer in some networks, we reduced the architecture of the SRGAN and added some layers to the FSRCNN. In the generator it was removed the 3 last layers and we reduced the number of downsample blocks by 2 in the discriminator in order to perform a 2x upscale. In the FSRCNN was added a deconvolution layaer at the end in order to perform 4x upscales.

During the tests we noticed that the VDSR trained with a Adam optimizer and without the skip layer have better results than the VDSR with the skip layer, as show in the results section. On the other hand there was no others modifications in the SRCNN, DRCN and RED-Net. During this work all networks were developed using the python's Keras library [38] with a TensorFlow backend [39].

4) Train and Tests: During the training process was used a batch of 32 images which each of those images were splited according to the process described in the data pre-processing subsection, creating an average of 192 images per batch. But in the case of the SRGAN this batch were reduced down to 5 images (30 sub images) since the GPU could not support the amount of images.

Each epoch used 1000 images and all networks were trained during 350 epochs. Each network took about 2 days to train each upscale factor. During our training steps we used both the Google cloud platform[4] equipped with an NVIDIA Tesla P100, along side with a dekstop with a Geforce 1060 6gb.

During the tests, all the test base were used. We used a multi-metric approach both the Jaccard distance and the PSNR were monitored during the training: an score would be only considered better then other if the Jaccard distance were bellow the 2.00 threshold and the PSNR higher then the old ones, this metric was used in order to prevent monochromatic images.

[4]https://console.cloud.google.com accessed in 23/06/2019

5) Post-processing: After the training and test step, the validation base was prepared to be used in the optical character recognition tests. In order to generate the validation images with 200DPI, 300DPI and 400DPI, we used a bicubic interpolation [18] in order to reduce the images by factors of 2x,1.5x,2x, after that we up sample all those images with a factor of 2x,2x,4x. Doing so to the 3 target resolution images. After up-scaling it using the networks and converted them to the gray scale color space, we used a white balance technique. It was also prepared a interpolated base, applying the bicubic interpolation [18] in the original image in order to have output with the same size as the neural network's output.

During our tests we also applied the Sauvola Threshold [40],with $Window = 60$ and $K = 0.01$, in order to evaluate the performance of the OCR engines in this scenario.

6) OCR Evaluation: During the tests, due to results of prior works [4], [41], we selected two popular optical character recognition engines: The Tesseract OCR (version 4.0) and the ABBYY fine reader engine (version 12.0). We used the standard configuration in the Tesseract while in the ABBYY we disabled the "Correct image resolution" option.

The input of both OCR engines was the post-processed images both the gray image and the binary image. In order to calculate the character error rate and the word error rate was used the Levenshtein distance [42]. During our the measure of CER special characters were ignored. This in WER case, both the special characters and multiple white spaces have been ignored

V. Results

A. OCR Results

After the experiments we measured the character error rate(CER), as shown in Table I, and the word error rate (WER), as shown in Table II. Both tables show that most of the super resolution network could decrease the error rate. Analyzing the tables we can affirm that:

- Comparing the 200 dpi gray scale generated output: Both the character and the word error rate get closer to the original one. Also some networks have gotten too close to the original rate. Thus is possible to assume that if the image is 2x downscaled, the SR image generated by one of this networks is close to the original image, using the OCR accuracy metric.
- Comparing the 300 DPI gray scale generated output: The decrease is higher than the 200 DPI image. Although in some cases the 400 DPI got lower results, we can assume that this is the best scenario due to the loss when compared with others resolutions. This is because some of the OCR engines are trained to use a 300 DPI image [43]
- Comparing the 400 dpi gray scale generated output: This case, has the lower error rate than the original, on the other hand some times the 300 DPI images have lower rates than this one, as expected. This was because in some point during the downsize there where some data loss, so

even the networks could not restore those details. But even with that data loss the results proven to be near to the original image

In the scenario where the Sauvola threshold [40] was applied to the image, is possible to notice a increase in the error rate when compared to the gray image. This because the OCR engines have some pre-processing step [43], doing so the performance of a gray image may be better than the performance of a binary image. Even though its possible to affirm that in this scenario the results have a behavior that close to the gray scale image.

TABLE I
THE CER FOR DIFFERENT SCENARIOS

Character Error Rate				
		200 DPI	300 DPI	400 DPI
Tesseract using binary	Bicubic [18]	63.62%	51.15%	52.34%
	DRCN [13]	54.09%	48.47%	45.99%
	RED-Net [14]	52.86%	47.58%	44.75%
	FSRCNN [8]	53.19%	49.37%	49.02%
	SRCNN [6]	53.75%	47.22%	43.48%
	SRGAN [15]	**46.65%**	**43.02%**	48.75%
	Modified VDSR	58.81%	46.32%	**42.97%**
	VDSR using the skip layer [7]	73.71%	60.23%	58.55%
	Original 200 DPI		50.45%	
	Low Resolution 100 DPI		68.58 %	
ABBYY using binary image	Bicubic [18]	70.92%	62.23%	59.17%
	DRCN [13]	62.24%	61.62%	52.33%
	RED-Net [14]	60.04%	60.83%	51.84%
	FSRCNN [8]	64.42%	62.15%	54.23%
	SRCNN [6]	62.65%	58.75%	51.64%
	SRGAN [15]	**56.61%**	**50.33%**	59.56%
	Modified VDSR	64.46%	58.67%	**50.49%**
	VDSR using the skip layer [7]	84.68%	70.36%	66.11%
	Original 200 DPI		58.50%	
	Low Resolution 100 DPI		75.34%	
Tesseract using gray image	Bicubic [18]	39.01%	30.71%	35.73%
	DRCN [13]	33.85%	28.69%	29.32%
	RED-Net [14]	32.25%	28.41%	30.37%
	FSRCNN [8]	34.40%	29.11%	31.54%
	SRCNN [6]	34.01%	28.52%	30.55%
	SRGAN [15]	**32.11%**	28.36%	**27.76%**
	Modified VDSR	36.70%	**28.09%**	29.57%
	VDSR using the skip layer [7]	70.97%	53.86%	49.90%
	Original 200 DPI		30.10%	
	Low Resolution 100 DPI		53.35%	
ABBYY using gray image	Bicubic [18]	36.16%	33.02%	36.36%
	DRCN [13]	40.36%	35.35%	31.40%
	RED-Net [14]	37.89%	34.09%	35.24%
	FSRCNN [8]	37.10%	35.06%	33.84%
	SRCNN [6]	38.56%	33.31%	**29.04%**
	SRGAN [15]	**35.13%**	**31.79%**	30.49%
	Modified VDSR	42.61%	32.67%	31.96%
	VDSR using the skip layer [7]	61.68%	51.72%	48.52%
	Original 200 DPI		27.82%	
	Low Resolution 100 DPI		49.23%	

B. Image Results

The results of the PSNR and the Jaccard distance are illustrated in table III. This results are calculated over the post-processed image both in the 200 DPI scenario.

TABLE II
THE WER FOR DIFFERENT SCENARIOS

Word Error Rate				
		200 DPI	300 DPI	400 DPI
Tesseract using binary	Bicubic [18]	65.72%	55.42%	56.52%
	DRCN [13]	56.09%	52.54%	50.68%
	RED-Net [14]	55.46%	51.77%	49.95%
	FSRCNN [8]	56.11%	52.97%	53.66%
	SRCNN [6]	56.48%	51.14%	48.34%
	SRGAN [15]	**50.41%**	**46.72%**	52.73%
	Modified VDSR	61.29%	50.58%	**48.02%**
	VDSR using the skip layer [7]	75.51%	63.66%	61.09%
	Original 200 DPI		53.47%	
	Low Resolution 100 DPI		71.23 %	
ABBYY using binary image	Bicubic [18]	71.87%	65.41%	63.88%
	DRCN [13]	64.82%	64.50%	57.25%
	RED-Net [14]	63.21%	64.55%	57.00%
	FSRCNN [8]	66.80%	65.58%	58.66%
	SRCNN [6]	65.29%	62.71%	56.90%
	SRGAN [15]	**60.32%**	**54.21%**	62.78%
	Modified VDSR	67.00%	61.96%	**55.64%**
	VDSR using the skip layer [7]	84.00%	72.40%	68.42%
	Original 200 DPI		62.12%	
	Low Resolution 100 DPI		76.40%	
Tesseract using gray image	Bicubic [18]	39.97%	33.28%	35.73%
	DRCN [13]	34.65%	31.10%	32.91%
	RED-Net [14]	**33.98%**	31.17%	33.29%
	FSRCNN [8]	35.89%	32.03%	35.04%
	SRCNN [6]	35.47%	31.37%	34.27%
	SRGAN [15]	34.27%	**30.48%**	**31.16%**
	Modified VDSR	37.64%	30.84%	33.35%
	VDSR using the skip layer [7]	71.52%	56.20%	51.63%
	Original 200 DPI		31.75%	
	Low Resolution 100 DPI		54.40%	
ABBYY using gray image	Bicubic [18]	40.69%	38.45%	42.74%
	DRCN [13]	44.07%	39.78%	36.42%
	RED-Net [14]	41.67%	38.62%	40.37%
	FSRCNN [8]	41.02%	39.33%	39.18%
	SRCNN [6]	42.84%	37.29%	**34.60%**
	SRGAN [15]	**39.18%**	**35.52%**	35.09%
	Modified VDSR	47.01%	37.15%	36.42%
	VDSR using the skip layer [7]	63.95%	54.71%	51.30%
	Original 200 DPI		32.66%	
	Low Resolution 100 DPI		52.45%	

As shown in table III the Jaccard distance is similar in all methods, having a small variance between them, that implies that the images are similar. However the PSNR values have a huge variance between them. This is probably because due to the amount of texture in the images background, as some of the images have more details in the background then others.

C. Relation between both results

Comparing Tables I, II and III, is shown that the higher PSNR do not implies in the lower error rate. This became evident in the DRCN case which have a good PSNR mean with a relatively low variance. However which was never the best result, but was one of the best results. Also the SRGAN is shown to have the best error rates but has a good PSNR. On the other hand the bicubic interpolation and the VDSR using the skip layer, were one of the worst error rate has the worst PSNR. So we may conclude that even if the PSNR do

not indicate if the error rate will decrease or increase, on the other hand a low PSNR may indicate a high error rate, as well as a high PSNR may indicate a low error rate.

TABLE III
PSNR AND JACCARD DISTANCE VALUES

Method	PSNR		Jaccard distance	
	Mean	Standard Derivation	Mean	Standard Derivation
Bicubic [18]	18,90	10,67	1,010976	$7{,}60e^{-6}$
DRCN [13]	21,04	13,39	1,010958	$7{,}27e^{-6}$
RED-Net [14]	20,94	14,01	1,010949	$7{,}28e^{-6}$
FSRCNN [8]	21,14	15,01	1,010972	$7{,}29e^{-6}$
SRCNN [6]	21,04	14,09	1,010975	$7{,}43e^{-6}$
SRGAN [15]	20,33	15,09	1,010947	$7{,}24e^{-6}$
VDSR	20,40	13,99	1,011012	$7{,}44e^{-6}$
VDSR using the skip layer [7]	19,38	5,17	1,011205	$8{,}03e^{-6}$

VI. Conclusion and Future work

In this paper, we explored the effectiveness of deep networks of images SR in a scenario with documents with a texture background. We first investigated and evaluated six existing deep networks under common daily situations. The objective was to evaluate the effectiveness of each model and to understand the best way to explore the benefits from these deep models for OCR task. Is also proven that the PSNR is lightly related to the OCR accuracy. Extensive experiments on SR with deep models demonstrate promissing results for OCR accuracy. In this work, we also proved that in different OCR applications the error rate is tends to decrease when the resolution of the image is increased in some scenarios only up to certain point.

Future work, we will evaluate the performance of recent networks such as ProSR [44], SRFBN [45], ESRGAN [46] and RCAN [47]. Also there is a high probability that we would develop a new deep architecture that is capable of handling the challenges of images SR for OCR applications.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, FACEPE and CNPq. The authors also would like to acknowledge the support and data provided by the company Callere Document Solutions.

References

[1] K. Umehara, J. Ota, and T. Ishida, "Application of super-resolution convolutional neural network for enhancing image resolution in chest ct," *Journal of Digital Imaging*, vol. 31, 2018.

[2] Magadomi. (2015) Waifu2x. [Online]. Available: http://waifu2x.udp.jp

[3] R. Pandey, S. Maiya, and A. Ramakrishnan, "A new approach for upscaling document images for improving their quality," in *2017 14th IEEE India Council International Conference (INDICON)*, 2017.

[4] A. Lat and C. Jawahar, "Enhancing ocr accuracy with super resolution," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018.

[5] H. Zhang, D. Liu, and Z. Xiong, "Cnn-based text image super-resolution tailored for ocr," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017.

[6] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, 2015.

[7] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[8] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European conference on computer vision*. Springer, 2016.

[9] F. Asad, A. Ul-Hasan, F. Shafait, and A. Dengel, "High performance ocr for camera-captured blurred documents with lstm networks," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, 2016.

[10] C. Chinara, N. Nath, S. Mishra, S. K. Sahoo, and F. A. Ali, "A novel approach to skew-detection and correction of english alphabets for ocr," in *2012 IEEE Student Conference on Research and Development (SCOReD)*. IEEE, 2012.

[11] Mande Shen and Hansheng Lei, "Improving ocr performance with background image elimination," in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2015.

[12] S. Susan and K. R. Devi, "Text area segmentation from document images by novel adaptive thresholding and template matching using texture cues," *Pattern Analysis and Applications*, pp. 1–13, 2019.

[13] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[14] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *arXiv preprint arXiv:1606.08921*, 2016.

[15] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of applied meteorology*, vol. 18, no. 8, 1979.

[17] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE transactions on image processing*, vol. 10, no. 10, 2001.

[18] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process*, pp. 1153–1160, 1981.

[19] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, "Deep learning for multiple-image super-resolution," *arXiv preprint arXiv:1903.00440*, 2019.

[20] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE transactions on image processing*, vol. 13, no. 10, 2004.

[21] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," 2015.

[22] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications*, no. 2, 2002.

[23] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan*, 2009.

[24] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, 1991.

[25] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, 1998.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014.

[27] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *arXiv preprint arXiv:1904.07523*, 2019.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[29] Y. Bengio, P. Simard, P. Frasconi *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, 1994.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[31] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, 2009.

[32] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.90

[34] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[36] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[38] F. Chollet *et al.*, "Keras," https://keras.io, 2015, accesed in 22/06/19.

[39] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016.

[40] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, 2000.

[41] C. Dong, X. Zhu, Y. Deng, C. C. Loy, and Y. Qiao, "Boosting optical character recognition: A super-resolution approach," *arXiv preprint arXiv:1506.02211*, 2015.

[42] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966.

[43] Google inc. (2019) Tesseract ocr. https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality. Accesed in 22/06/19.

[44] W. Yifan, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in *CVPR Workshops*, June 2018.

[45] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[46] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[47] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.