

UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E
SISTEMAS

João Batista Pacheco Junior

**USO DE REDES NEURAIAS CONVOLUCIONAIS NA SEGMENTAÇÃO DE VIAS
URBANAS ASFALTADAS EM IMAGENS DE SATÉLITE RGB: Estudo de caso em
São Luís-MA**

São Luís – MA

2019

João Batista Pacheco Junior

**USO DE REDES NEURAIAS CONVOLUCIONAIS NA SEGMENTAÇÃO DE VIAS
URBANAS ASFALTADAS EM IMAGENS DE SATÉLITE RGB: Estudo de caso em
São Luís-MA**

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Computação e Sistemas como requisito parcial para obtenção do título de Mestre em Engenharia de Computação e Sistemas com área de concentração em Tecnologia da Informação.

Orientador: Prof. Me. Henrique Mariano Costa do Amaral

Co-orientador: Prof. Me. Antônio Fernando Lavareda Jacob Júnior

São Luís – MA

2019

Pacheco Junior, João Batista.

Uso de Redes Neurais Convolucionais na segmentação de vias urbanas asfaltadas em imagens de satélite RGB: estudo de caso em São Luís – MA / João Batista Pacheco Junior. – São Luís, 2019.

60 f.

Dissertação (Mestrado) – Curso de Engenharia de Computação e Sistemas, Universidade Estadual do Maranhão, 2019.

Orientador: Henrique Mariano Costa do Amaral.

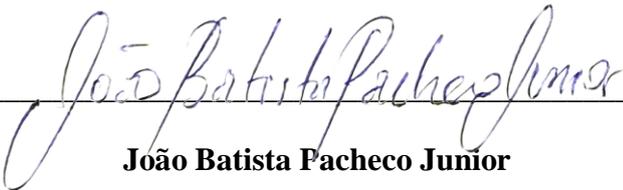
1. Geoprocessamento. 2. Imagem RGB. 3. Segmentação de imagens. 4. Rede Neural Convolucional. I. Título.

CDU: 004.8.032.26:528.854(812.1)

João Batista Pacheco Junior

**USO DE REDES NEURAIAS CONVOLUCIONAIS NA SEGMENTAÇÃO DE VIAS
URBANAS ASFALTADAS EM IMAGENS DE SATÉLITE RGB: Estudo de caso em
São Luís-MA**

Aprovada em 09/10/2019.



João Batista Pacheco Junior

Orientando



Prof. Me. Henrique Mariano Costa do Amaral

Orientador



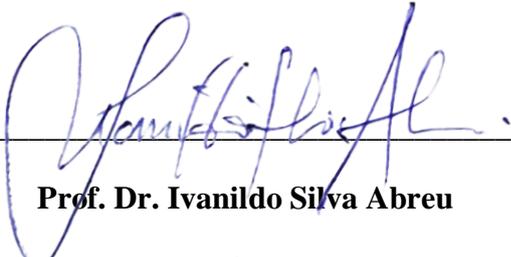
Prof. Me. Antônio Fernando Lavareda Jacob Junior

Co-orientador



Prof. Dr. Luís Carlos Costa Fonseca

Examinador Interno



Prof. Dr. Ivanildo Silva Abreu

Examinador Externo

Agradecimentos

A Deus, pela força, benção e esperança renovada de que o amanhã poderá sempre ser melhor.

Aos meus pais, pilares essenciais de minhas conquistas.

Ao meu orientador, Prof. Dr. Henrique Mariano Costa do Amaral, e co-orientador Prof. Antônio Fernando Lavareda Jacob Júnior, pela prestatividade, auxílio e orientação sempre muito construtivos.

A todos os demais professores, em especial aos do Programa de Pós-graduação em Engenharia de Computação e Sistemas, pelos ensinamentos que contribuíram direta ou indiretamente para a realização deste trabalho.

A Jeff Heaton, Ph.D./Washington University, pela boa vontade dos e-mails trocados. Suas valiosas dicas não teriam tornado esta ideia viável.

A Mayana Almeida e meu pai, João Batista, por unir forças no desenho das centenas de ROIs.

A todos aqueles que contribuíram de qualquer forma para o sucesso deste trabalho.

“A experiência é uma escola onde são caras as lições, mas em nenhuma outra os tolos podem aprender”.

Benjamin Franklin

Resumo

É frequente usar imagens de satélite em atividades de geoprocessamento e editar geometrias vetoriais relativas a cada aplicação, sendo o desenho manual de vias uma atividade bastante frequente. Com a crescente demanda por desenhos vetoriais em SIGs, em especial a verificação e edição de vias terrestres, torna-se conveniente o desenvolvimento de uma ferramenta computacional que, com a ajuda da inteligência artificial, automatize o que for possível nesse aspecto, já que tal edição manual depende dos limites da agilidade do usuário.

Para testar a viabilidade desta proposta, uma base de dados composta por 600 imagens RGB é apresentada a três diferentes Redes Neurais Convolucionais (CNNs) projetadas para segmentação de objetos e o desempenho de cada uma foi avaliado e comparado tendo-se em vista acurácia, *IoU* e tempo de segmentação.

Nas condições do experimento, a CNN SegNet se mostrou mais apropriada à segmentação de vias urbanas asfaltadas em imagens RGB em aplicações de precisão não sensíveis a descontinuidades, logrando acurácia média de 87,12% e *IoU* médio de 71,93%, ao passo que a Dilated Convolutions alcançou os valores de 85,27% e 61,27% e a U-Net 78,09% e 64,91%, respectivamente. Alternativamente, Dilated Convolutions mostrou maior afinidade nesta tarefa em aplicações que demandem um processamento de alta frequência e que não sofram tanta interferência dos falsos positivos, por conciliar boa acurácia com um tempo médio bastante reduzido de 31,2 ms para imagens de 256×256 pixels e 125 ms para as de 1024×1024 pixels, contra os respectivos 78,1 e 546,9 da U-Net e 93,7 e 687,5 ms da SegNet.

Palavras-chave: Geoprocessamento, imagem RGB, segmentação de imagens, Rede Neural Convolucional.

Abstract

Satellite imagery is often used in geoprocessing activities and editing vector geometries for each application, with manual road design being a very frequent activity. With the growing demand for GIS vector drawings, especially the verification and editing of land roads, it is convenient to develop a computational tool that, with the help of artificial intelligence, automates what is possible in this respect, since such manual editing depends on the limits of user agility.

To test the feasibility of this proposal, a database composed of 600 RGB images is presented to three different Convolutional Neural Networks (CNNs) designed for segmentation of objects and their performance was evaluated and compared for accuracy. and segmentation time.

Under the conditions of the experiment, CNN SegNet was more appropriate for segmentation of asphalted urban roads in RGB images in non-discontinuous precision applications, achieving an average accuracy of 87.12% and an average IoU of 71.93%, whereas Dilated Convolutions reached 85.27% and 61.27% and U-Net 78.09% and 64.91%, respectively. Alternatively, Dilated Convolutions has shown greater affinity for this task in applications that require high frequency processing and not so much interference from false positives, because it combines good accuracy with a greatly reduced average time of 31.2 ms for 256×256 pixel images. and 125 ms for 1024×1024 pixels, against U-Net's 78.1 and 546.9 and SegNet's 93.7 and 687.5 ms.

Keywords: Geoprocessing, RGB image, image segmentation, Convolutional Neural Network.

Lista de figuras

Figura 1. a) Uma imagem 256×256 com 256 níveis de cinza. b) Representação gráfica desta imagem. (13).....	6
Figura 2. Absorção das luzes vermelha, verde e azul pelos sensores do olho humano como função do comprimento de onda (8).....	7
Figura 3. a) Cores primárias da luz e suas combinações aditivas. b) Cores primárias de pigmentos e suas combinações subtrativas. (8).....	7
Figura 4. O subespaço de cores RGB, com R, G e B interpolados no intervalo $0, 1$. No exemplo, a cor indicada está associada ao <i>pixel</i> de coordenadas $x = 287$ e $y = 134$ da imagem 512×512 <i>pixels</i> . A diagonal principal do cubo representa a reta dos níveis de cinza, que vão do preto $0, 0, 0$ ao branco $1, 1, 1$	8
Figura 5. Ilustração do processo de convolução em imagens como um filtro espacial linear....	9
Figura 6. Esquema de convolução de um filtro <i>sharpening</i> (em azul) sobre uma imagem 4×4 (em amarelo) com <i>zero-padding</i> igual a 1, no qual o valor do <i>pixel</i> é calculado através da equação (6).	10
Figura 7. Modelo de neurônio (17), no qual eventualmente pode ser adicionado um valor definido pelo projetista chamado <i>bias</i> , para controlar o comportamento de certos modelos neurais.....	12
Figura 8. Ilustração resumida de uma CNN classificando uma imagem como “cidade” ou “praia”.....	14
Figura 9. Imagem de ressonância magnética de um tumor cerebral segmentado e classificado por uma CNN, dividido por classes: tumor em crescimento (vermelho), tumor de tamanho estabilizado (amarelo), azul (necrose), verde (edema) (27).	14
Figura 10. Para cada área da imagem, a convolução do filtro alimenta um neurônio por camada do volume de saída. Neste caso, é representada uma coluna de profundidade de 5 neurônios do volume de saída, na qual cada neurônio é resultado da convolução do filtro associado sobre uma área específica.....	16

Figura 11. A convolução dos N filtros sobre a imagem gera N camadas resultantes, que juntas formam o volume de saída.....	18
Figura 12. a) <i>Pooling</i> através da função MAX, que preserva o maior valor dentro da máscara. b) <i>Pooling</i> através da função AVERAGE, que mantém a média dos valores interiores à máscara. c) Exemplo de <i>downsampling</i> de uma imagem RGB após <i>max pooling</i> da imagem. A quantidade de canais ($D = 3$) foi preservada.....	19
Figura 13. Arquitetura da AlexNet, uma CNN comumente usada para classificação de imagens (31).....	20
Figura 14. Esquema geral de uma segmentação semântica (38).....	21
Figura 15. Arquitetura U-Net. Cada caixa azul corresponde a um mapa de características multicanal. O número de canais é indicado na parte superior da caixa. As dimensões da imagem ao longo do processo são indicadas no canto inferior esquerdo da caixa. Caixas brancas representam mapas de características copiados. As setas denotam as diferentes operações ao longo do processo (32).	22
Figura 16. Representação do decodificador SegNet. Os índices são utilizados para montar o próprio mapa de características a ser convolvido por um filtro decodificador treinável (33).....	23
Figura 17. A dilatação suporta uma expansão exponencial do campo receptivo sem perda de resolução. a) Filtro para convolução 1-dilatada ($k = 3$) (a), 2-dilatada ($k = 7$) (b) e 4-dilatada ($k = 15$) (c) (34).....	24
Figura 18. Etapas da pesquisa proposta em sequência.	26
Figura 19. Percentual de reflectância das classes mais comuns em paisagens urbanas (41). É nítida sua diferenciação no espectro infravermelho, não presente em grande parte dos geoserviços gratuitos.	27
Figura 20. Edição de uma ROI no <i>Matlab Image Labeler</i> . Na figura ao centro, os <i>pixels</i> correspondentes à classe 1 (vias) estão em azul e os à classe 2 (outros) estão em laranja.....	28
Figura 21. a) Os diversos elementos poluidores da imagem destacados em retângulos: sedimentos arenosos sobre a pista (azul), arborização aleatória (vermelho), sombras (amarelo), tons de concreto em consertos de erosões no asfalto (verde) e marca d'água do Google (canto inferior direito). b) imagem do item a) com rua ROI desenhada sem considerar os referidos elementos poluidores.	29
Figura 22. Diagrama de Venn para as variáveis preditivas.	31
Figura 23. Pior (a) e melhor (b) caso de teste para a U-Net referente à classe “Vias”.....	33
Figura 24. Pior (a) e melhor (b) caso de teste para a SegNet referente à classe “Vias”.....	34

Figura 25. Pior (a) e melhor (b) caso de teste para a Dilated Convolutions referente à classe “Vias”.....	36
Figura 26. Comparativo do tempo médio (segundos) de segmentação de uma imagem aleatória da base de testes por cada CNN.....	37

Lista de abreviaturas e siglas

ANN	<i>Artificial Neural Network</i>
CNN	<i>Convolutional Neural Network</i>
CPU	<i>Central Processing Unit</i>
DDR4	<i>Double Data Rate 4</i>
FCN	<i>Fully Convolutional Networks</i>
FN	Falso Negativo
FP	Falso Positivo
GDDR5	<i>Graphics Double Data Rate 5</i>
GPU	<i>Graphic Processing Unit</i>
IoU	<i>Intersection over Union</i>
KNN	<i>K-Nearest Neighbor</i>
MLP	<i>Multi-Layer Perceptron</i>
NVMe	<i>Non-Volatile Memory express</i>
PDI	Processamento Digital de Imagens
ReLU	<i>Rectified Linear Unit</i>
RGB	<i>Red-Green-Blue</i>
ROI	<i>Region of Interest</i>
SDRAM	<i>Synchronous Dynamic Random-Access Memory</i>
SIANN	<i>Shift-Invariant Artificial Neural Network</i>
SIG	Sistema de Informação Geográfica
SOM	<i>Self-Organized Map</i>
SSD	<i>Solid State Drive</i>
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

Lista de tabelas

Tabela 1. Arquitetura da rede de contexto. A rede processa C mapas de características agregando informação contextual em escalas progressivamente crescentes sem perder a resolução (34).	25
Tabela 2. Comparativo geral entre as CNNs U-Net, SegNet e Dilated Convolutions.	37

Sumário

Agradecimentos	i
Resumo	iii
Abstract	iv
Lista de figuras	v
Lista de abreviaturas e siglas	viii
Lista de tabelas	ix
Sumário	x
Capítulo 1 Introdução	1
1.1 Justificativa	1
1.2 Trabalhos relacionados	2
1.3 Objetivos.....	3
1.3.1 OBJETIVO GERAL	3
1.3.2 OBJETIVOS ESPECÍFICOS.....	3
1.4 Organização do trabalho	3
Capítulo 2 Fundamentação teórica	5
2.1 Processamento Digital de Imagens	5
2.1.1 IMAGENS DIGITAIS.....	6
2.1.2 O ESPAÇO DE CORES RGB	6
2.1.3 CONVOLUÇÃO E FILTRAGEM.....	8
2.2 Redes Neurais Artificiais	11
2.2.1 MODELO DE UM NEURÔNIO E ARQUITETURAS DE REDES NEURAIAS.....	11
2.2.2 APRENDIZAGEM SUPERVISIONADA E NÃO-SUPERVISIONADA	12
2.3 Redes Neurais Convolucionais	13
2.3.1 HIPERPARÂMETROS	15
2.3.1.1 Passo	15
2.3.1.2 Profundidade	15
2.3.1.3 Tamanho da camada	16
2.3.1.4 Taxa de aprendizagem.....	16
2.3.1.5 Função de ativação	17
2.3.2 CAMADA CONVOLUCIONAL	17
2.3.3 CAMADA DE POOLING	18
2.3.4 ARQUITETURAS PARA SEGMENTAÇÃO DE IMAGENS	20
2.3.4.1 U-Net	21

2.3.4.2 Segnet	22
2.3.4.3 Dilated Convolutions.....	23
Capítulo 3 Materiais e métodos.....	26
3.1 Visão geral.....	26
3.1.1 AQUISIÇÃO DE IMAGENS	27
3.2 Base de Dados e pré-treino	28
3.3 Treinamento da Rede Neural	29
3.4 Métricas de avaliação	30
Capítulo 4 Resultados e Discussões.....	32
4.1 Avaliação da U-Net	32
4.2 Avaliação da SegNet.....	33
4.3 Avaliação da Dilated Convolutions	35
4.4 Comparação entre U-Net, SegNet e Dilated Convolutions	36
Capítulo 5 Considerações Finais	39
Referências	41

Capítulo 1

Introdução

1.1 Justificativa

O uso cada vez mais frequente dos SIGs para diversas atividades ligadas ao geoprocessamento requer um nível cada vez maior de qualidade das informações. O georreferenciamento dos objetos às imagens de satélite dos territórios requer, na maior parte dos casos, a detecção visual por operador humano e a vetorização manual das geometrias que garantem tal georreferenciamento. Este processo demanda considerável tempo e mão-de-obra, dada a enorme quantidade de elementos presentes nas imagens.

Um dos processos feitos manualmente em diversos trabalhos de geoprocessamento é a incorporação de novas vias terrestres (isto é, ruas, avenidas, rodovias, estradas, becos, etc.) nos bancos de dados geográficos, à medida que imagens de satélite mais atualizadas do território são adquiridas. Continuamente, surgem novas ocupações urbanas (bairros, aglomerados subnormais, áreas urbanas isoladas, etc.) e rurais (povoados, lugarejos, projetos de assentamento, etc.), para as quais precisam ser desenhadas as geometrias vetoriais específicas para caracterização dos mapas (em especial, arruamentos) e acomodação dos dados relevantes associados. Tal atividade requer que o usuário analise diversas áreas do território coberto pela imagem, e atualize a geometria das vias de acordo com as mudanças detectadas, condicionando tal processo aos limites da agilidade humana.

Em face do contexto apresentado, é indiscutível o aumento de produtividade tendo-se em mãos ferramentas que automatizem o que for possível de tal processo, utilizando-se os recursos disponíveis, como imagens aéreas do território e inteligência artificial.

1.2 Trabalhos relacionados

A situação apresentada motiva diversos trabalhos na literatura científica que propõem detecção de vias em imagens de sensoriamento remoto. Dentre eles, podemos citar Pinho et al. (1), que comparam diferentes métodos de classificação de imagens urbanas de alta resolução espacial. Utilizando imagens do satélite IKONOS, os autores concluíram que os métodos de classificação orientados a objeto são mais adequados que os métodos usuais para classificar *pixels* mistos, isto é, *pixels* que pertencem a mais de uma classe de objetos se avaliados por algoritmos tradicionais de classificação. A utilização de classes e objetos possibilita a inserção de elementos tradicionais de interpretação visual na forma de descritores, como cor, forma, tamanho, textura, padrão e contexto.

Além deste, Simões (2) utiliza ANNs para segmentar imagens através da classificação de cores, almejando maximizar o desempenho do classificador determinando uma relação entre os atributos das imagens e os diferentes casos de estudo.

Nesta linha, o trabalho de Venturieri (3) também utiliza algoritmos de retropropagação para melhorar a taxa de sucesso do uso de ANNs em caracterização do uso de solos, utilizando atributos de cor e textura.

Já Nóbrega (4) utiliza técnicas de detecção de vias através de classificação orientada a objetos em imagens de alta resolução espacial. Em virtude da heterogeneidade de elementos presentes nas imagens, foi feita uma identificação e segregação das diferentes feições em classes específicas, tendo em vista as características espectrais, geométricas e contextuais de cada uma.

Ainda no escopo da classificação de imagens, Pinho et al. (5) utilizam o algoritmo C4.5 como uma melhoria no processo de seleção de atributos de classificação de imagens, obtendo resultados similares, com tempo bastante reduzido. A técnica consistiu em um aprimoramento da classificação orientada a objetos.

Por fim, Doucette et al (6) também realizaram um outro trabalho notável na detecção de vias em imagens de satélite multiespectrais, através um método não supervisionado de classificação utilizando SOMs.

As Redes Neurais Convolucionais (CNNs) são, atualmente, consideradas o estado-da-arte em *deep learning* aplicado ao processamento de imagens e reconhecimento de padrões (7).

Tal desempenho pode eventualmente viabilizar aplicações em tempo real na edição de feições geométricas.

1.3 Objetivos

1.3.1 OBJETIVO GERAL

Este estudo visa discutir e avaliar a aplicação de diferentes arquiteturas de CNNs na segmentação da malha de vias asfaltadas a partir de uma imagem aérea de determinado território urbano, oriunda de geoserviço.

1.3.2 OBJETIVOS ESPECÍFICOS

Para alcançar este objetivo, é especificamente necessário:

- 1) Montar uma base de dados a partir da obtenção de imagens aéreas coloridas da área de estudo e da definição de suas respectivas *labels*;
- 2) Treinar CNNs de arquiteturas U-Net, SegNet e Dilated Convolutions para segmentar a região de interesse;
- 3) Avaliar e comparar o desempenho das três arquiteturas na segmentação da região de interesse.

1.4 Organização do trabalho

O Capítulo 2 apresenta a fundamentação teórica deste trabalho, discutindo os principais tópicos em Processamento Digital de Imagens, incluindo definição, breve histórico, espaço de cores RGB e tópicos selecionados sobre convolução e filtragem. Inclui, ainda, os conceitos básicos em ANNs, como modelo neuronal e tipos de aprendizagem. A partir disto, por fim, este capítulo fala sobre as CNNs, seus principais hiperparâmetros e sobre as arquiteturas que serão utilizadas neste estudo.

Em seguida, no Capítulo 3, são apresentados os recursos utilizados e as etapas da metodologia empregada para alcançar os objetivos deste trabalho: aquisição das imagens de

estudo, montagem da base de dados, preparação e treinamento das redes neurais e definição das métricas de avaliação.

Finalmente, no Capítulo 4, são apresentados os resultados do experimento e a avaliação do desempenho das redes neurais, tendo em vista a base de dados e as métricas definidas no capítulo anterior.

Capítulo 2

Fundamentação teórica

2.1 Processamento Digital de Imagens

O *Processamento Digital de Imagens* (PDI) é um campo da Ciência da Computação que estuda o uso do computador digital para o tratamento e aplicação de imagens digitais (8) e suas aplicações são amplamente variadas.

Nas aplicações médicas, podemos citar Reis (9), que utiliza o modelo de Ising na detecção automatizada da próstata e classificação de lesões em imagens de exames de ressonância magnética. Um outro exemplo é Braz Junior et al (10), que propõem um método de detecção de câncer de mama em imagens de mamografia com precisão relativamente alta.

No campo da visão computacional, são alguns dos inúmeros exemplos o estudo de Chatrath et al (11) que apresenta melhorias técnicas para a detecção e rastreamento ágil de faces humanas em vídeos de vigilância; e o trabalho de George et al (12) sobre a detecção de sorrisos em imagens estáticas utilizando o algoritmo KNN.

De modo geral, a lista de aplicações do PDI é grande e diversificada. Muitas são encontradas, também, na robótica (visão robótica), no sensoriamento remoto (classificação e uso do solo, detecção de recursos, avaliação de evolução de desmatamento, etc.), na inspeção automática em indústrias, em aplicativos móveis de fotografia, entre outras (13).

Nas seções a seguir, serão discutidos alguns tópicos em PDI relevantes para a compreensão do tema proposto neste estudo.

2.1.1 IMAGENS DIGITAIS

Uma *imagem digital* pode ser entendida matematicamente como uma função discreta $f(x, y)$ na qual x e y são coordenadas¹ espaciais da imagem, juntos correspondem a um *pixel* e seu valor resultante corresponde à *intensidade*² da imagem (8).

Pixels de imagens digitais podem assumir uma quantidade N de valores limitada pelo número b de bits utilizados para armazená-los (14). Matematicamente,

$$N = 2^b. \quad (1)$$

Em imagens monocromáticas o mais usual são 8 bits, o que permite representar 256 tons de cinza.

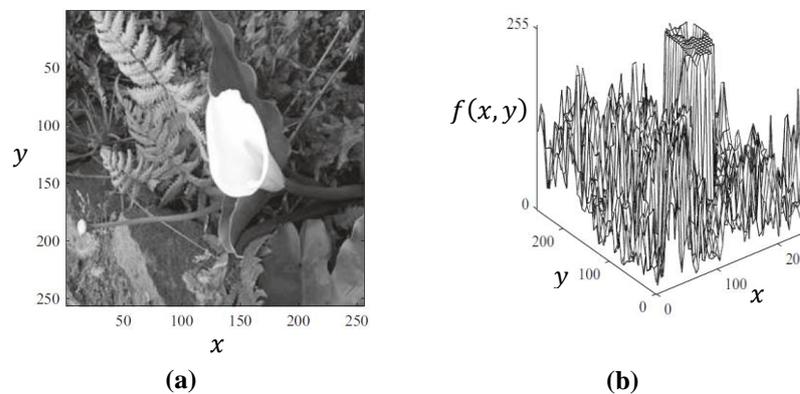


Figura 1. a) Uma imagem 256×256 com 256 níveis de cinza. b) Representação gráfica desta imagem. (13)

Além das imagens monocromáticas, existem as coloridas. A cor é um atributo descritivo eficiente, que comumente torna identificar e extrair um objeto de uma cena tarefas relativamente simples. Neste estudo, as imagens utilizadas serão coloridas.

2.1.2 O ESPAÇO DE CORES RGB

Gonzalez et al (8) aludem a estudos que mostram que os sensores do olho humano se agrupam em basicamente três grupos: os sensíveis ao vermelho (*red*, cerca de 65%), os

¹ As coordenadas de uma imagem digital também são valores discretos (8).

² A intensidade também é usualmente referida como *nível de cinza* (51).

sensíveis ao verde (*green*, cerca de 33%) e os sensíveis ao azul (*blue*, cerca de 2%), sendo estes os mais sensíveis. Esta ideia é ilustrada na Figura 2.

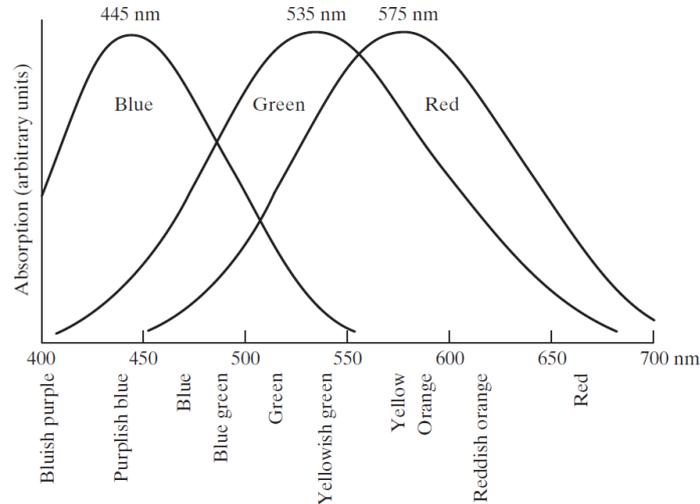


Figura 2. Absorção das luzes vermelha, verde e azul pelos sensores do olho humano como função do comprimento de onda (8).

As três referidas cores são ditas *cores primárias da luz* pois, sob certas condições, podem compor³ novas cores ao serem misturadas (Figura 3-a). Por outro lado, ciano (cyan), magenta (magenta) e amarelo (yellow) são as cores que possuem a propriedade de absorver⁴ uma das três cores primárias da luz enquanto refletem todas as outras, e são denominadas *cores secundárias de pigmentos* (Figura 3-b) (8).

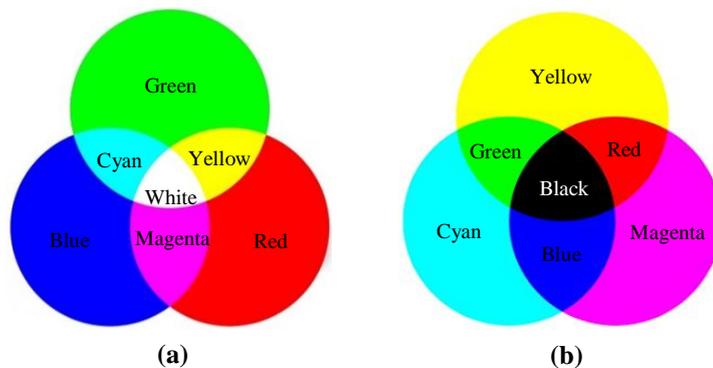


Figura 3. a) Cores primárias da luz e suas combinações aditivas. b) Cores primárias de pigmentos e suas combinações subtrativas. (8)

³ Por formarem novas cores a partir de uma mistura, dizemos que as cores primárias da luz possuem *princípio aditivo* (8) (13) (14).

⁴ Analogamente, as cores primárias de pigmentos possuem *princípio subtrativo* (8) (13) (14).

O modelo de cores RGB é uma relação de associação de três intensidades de cor primária da luz (ou seja, r , g e b) a uma cor composta f . Como os objetos deste estudo são imagens bidimensionais, cada cor f está associada à coordenada (x, y) de um *pixel*. Portanto, matematicamente,

$$f(x, y) = (r, g, b). \quad (2)$$

Desta forma, é possível representar este modelo de cores por um sistema cartesiano tridimensional, no qual cada dimensão expressa a intensidade de uma cor primária da luz e cada ponto está associado a um *pixel* da imagem, conforme ilustrado na Figura 4.

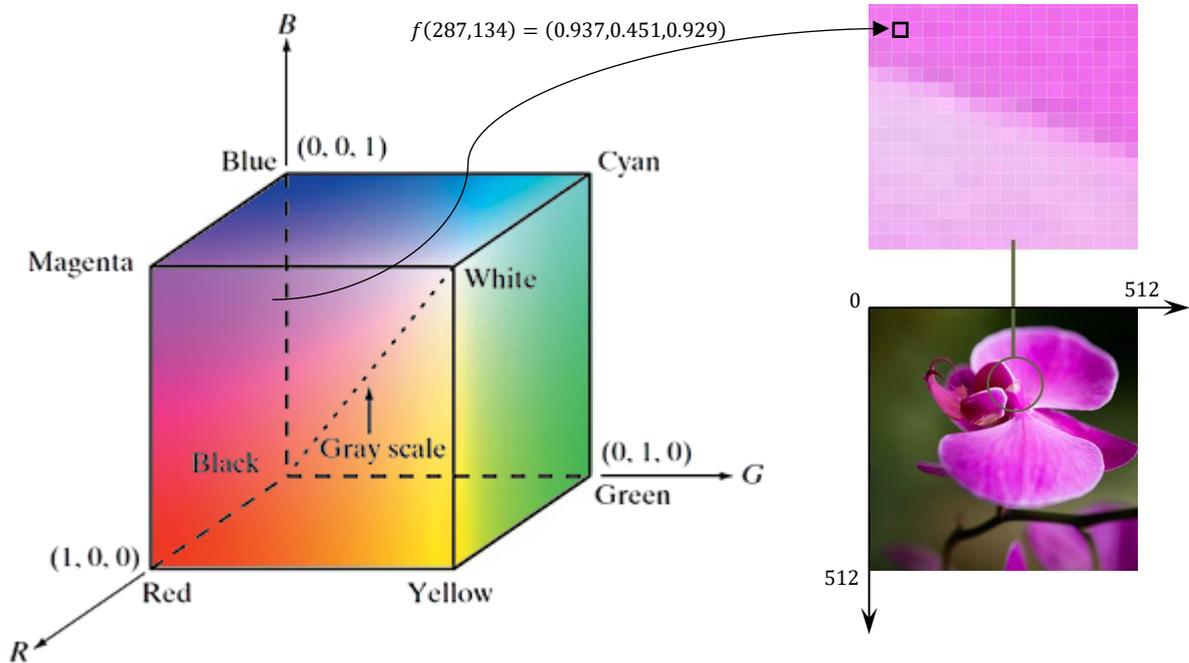


Figura 4. O subespaço de cores RGB, com R, G e B interpolados no intervalo $[0, 1]$. No exemplo, a cor indicada está associada ao *pixel* de coordenadas $x = 287$ e $y = 134$ da imagem 512×512 *pixels*. A diagonal principal do cubo representa a reta dos níveis de cinza, que vão do preto $(0, 0, 0)$ ao branco $(1, 1, 1)$.

2.1.3 CONVOLUÇÃO E FILTRAGEM

Seja f e h funções contínuas em \mathbb{R} , onde h corresponde a um sinal de entrada, que se desloca ao longo do tempo x . A *convolução* de f em resposta ao sinal h resulta em uma função g , que é dada por

$$g(x) = f(x) * h(x) = \int_{-\infty}^{+\infty} f(k) \cdot h(x - k) dk. \quad (3)$$

Trata-se, portanto, de um operador linear que expressa a sucessiva soma do produto dessas duas funções ao longo da região subentendida pela superposição delas em função do deslocamento pela abcissa correspondente (15). Posto isto, a convolução discreta é dada por

$$g(x) = f(x) * h(x) = \sum_{k=-F}^F f(k) \cdot h(x - k), \quad (4)$$

com $F \in \mathbb{N}$ correspondendo ao comprimento do filtro.

Analogamente, para casos nos quais o sinal é bidimensional (como é o caso das imagens), sua representação se dá por uma função de duas variáveis e, portanto, tem-se

$$g(x, y) = f(x, y) * h(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(u, v) \cdot h(x - u, y - v) dudv, \quad (5)$$

para a qual a versão discreta é dada por

$$g(x, y) = f(x, y) * h(x, y) = \sum_{u=-F_1}^{F_1} \sum_{v=-F_2}^{F_2} f(u, v) \cdot h(x - u, y - v). \quad (6)$$

onde $F_1, F_2 \in \mathbb{N}$ representam as dimensões do filtro espacial.

A equação (6) é frequentemente utilizada no PDI como um *filtro espacial linear*⁵ (Figura 5) na suavização e realce de detalhes e redução de ruídos (16). A Figura 6 ilustra um filtro do tipo *sharpening* de tamanho $F = 3$ convoluindo sobre uma imagem 4×4 pixels.

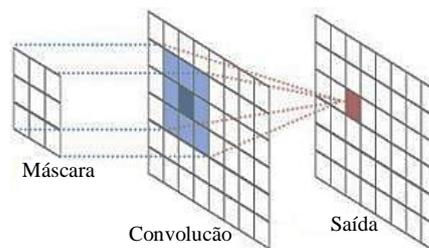


Figura 5. Ilustração do processo de convolução em imagens como um filtro espacial linear.

⁵ Filtros espaciais lineares também são comumente chamados de máscara, *kernel*, *template* ou janela (8).

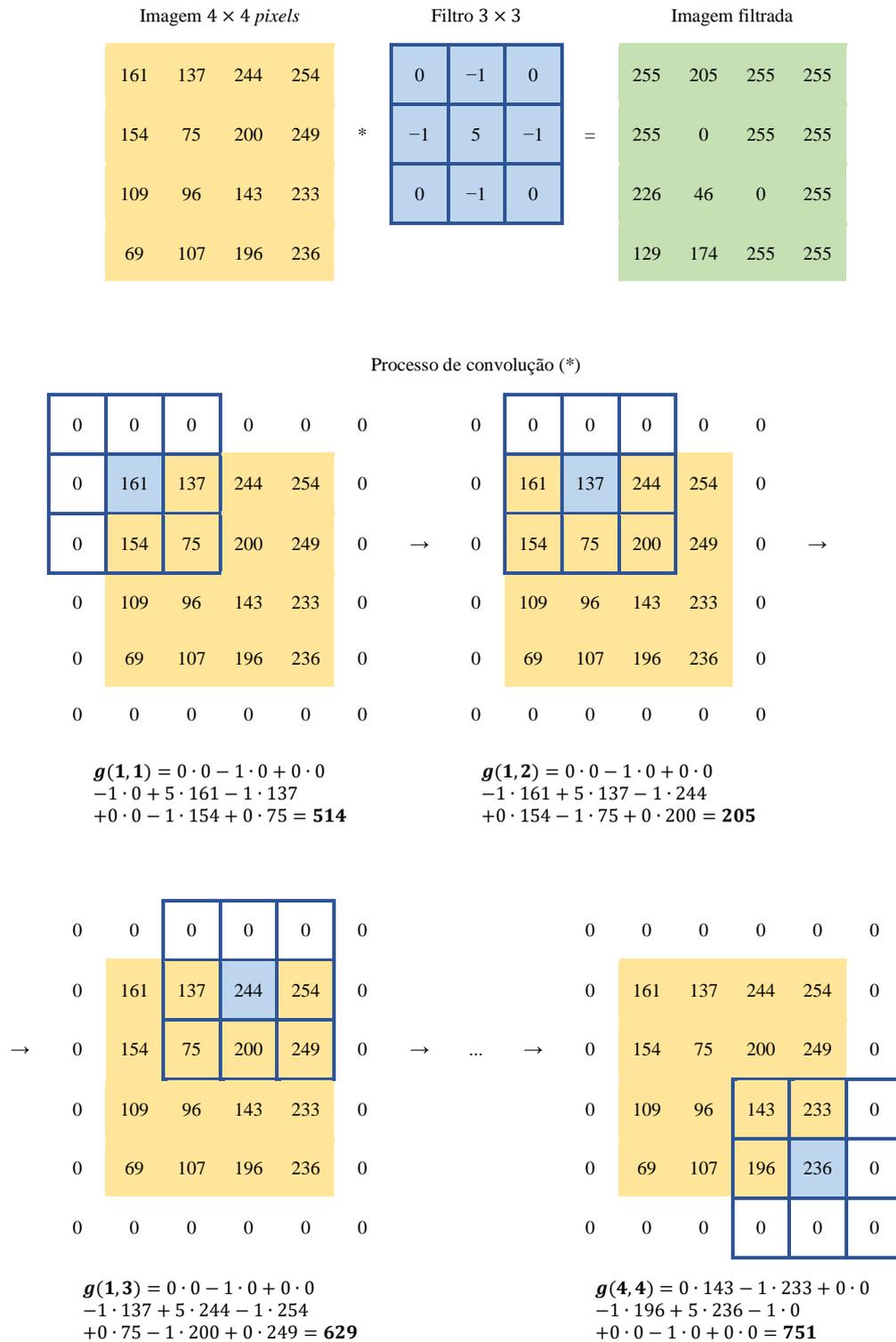


Figura 6. Esquema de convolução de um filtro *sharpening* (em azul) sobre uma imagem 4×4 (em amarelo) com *zero-padding*⁶ igual a 1, no qual o valor do *pixel* é calculado através da equação (6).

⁶ *Zero-padding* é o preenchimento, usualmente com zeros, de um sinal (neste caso, uma imagem) de modo que haja células suficientes a serem convoluídas pelo *kernel* em qualquer *pixel* (22). O número p de filas de *zero-padding* é dado por $p = \frac{n-1}{2}$, onde n é a dimensão do *kernel*.

2.2 Redes Neurais Artificiais

De acordo com Haykin (17), uma *Rede Neural Artificial* ou ANN (acrônimo do inglês *Artificial Neural Network*) é uma estrutura de dados em forma de processador com alto grau de distribuição paralela, constituído de unidades de processamento simples projetadas para armazenar conhecimento experimental e disponibilizá-lo para o uso. São inspiradas no funcionamento do cérebro humano, que possui basicamente duas premissas (17):

- i. *O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.*
- ii. *Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.*

ANNs são, portanto, um algoritmo capaz de gerar uma base de conhecimento a partir de informações a ele fornecidas e de resolver certos tipos de problemas aplicando esse conhecimento obtido. Tal ferramenta possibilitou a concretização de trabalhos como o de Dawson e Wilby (18), que apresentaram uma arquitetura de ANN capaz de modelar o comportamento de chuvas e prever cheias; e o de Khan et al (19) que desenvolveram uma ANN capaz de classificar cânceres para categorias específicas de diagnóstico com base em suas assinaturas de expressão genética, contornando dilemas diagnósticos na prática clínica e identificando os genes mais relevantes para a classificação.

Os tópicos mais relevantes em ANN para a compreensão do tema proposto neste estudo são percorridos nas seções seguintes.

2.2.1 MODELO DE UM NEURÔNIO E ARQUITETURAS DE REDES NEURAIS

Uma ANN é constituída de unidades de processamento denominadas *neurônios*, cujo funcionamento é uma abstração do funcionamento do neurônio humano.

Este neurônio recebe os sinais de entrada através de conexões chamadas *dendritos*. Estes sinais são potencializados de acordo com o peso de cada conexão, denominado *peso sináptico*. Em seguida, estes sinais são somados no *corpo neuronal* para então, finalmente, serem filtrados no *axônio* por uma *função de ativação*, gerando um sinal de saída (20). Este processo está ilustrado na Figura 7.

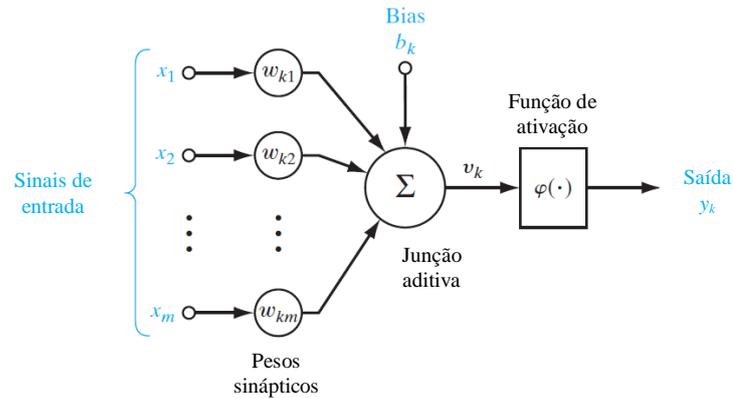


Figura 7. Modelo de neurônio (17), no qual eventualmente pode ser adicionado um valor definido pelo projetista chamado *bias*, para controlar o comportamento de certos modelos neurais.

Após receber um sinal de entrada x_i , o neurônio artificial fornece somatório v_k , que é obtido a partir do produto escalar entre o vetor $X = (x_1, x_2, \dots, x_m)$ correspondente a uma ou mais entradas específicas e o vetor $W_k = (w_{k1}, w_{k2}, \dots, w_{km})$ dos pesos sinápticos. Matematicamente,

$$v_k = \sum_{i=0}^n w_{km} x_m, \quad (7)$$

onde k é o índice de cada neurônio e j o índice de cada entrada e peso, com $k, m \in \mathbb{N}^*$.

Após a junção aditiva, este valor é eventualmente somado a um bias de controle b_k e passa por uma função de ativação φ , gerando a saída y_k . Podemos escrever este processo (17) na forma da equação

$$y_k = \varphi(v_k + b_k). \quad (8)$$

2.2.2 APRENDIZAGEM SUPERVISIONADA E NÃO-SUPERVISIONADA

Conforme mencionado, as ANNs adquirem conhecimento através de um processo de aprendizagem, que é variado e depende da arquitetura e funcionamento da rede.

Classificamos como *aprendizagem supervisionada* todo treinamento feito com base no fornecimento de prévio de uma coleção de pares de entradas e saídas à ANN, para que esta ajuste seus parâmetros até obter a melhor relação possível (isto é, uma *função-objetivo* ótima) entre a entrada e a saída de cada par (21). A cada iteração t , o valor do erro $e(t)$ a ela associado

é avaliado pela ANN através da diferença entre o valor $d(t)$ desejado e a saída $y(t)$ calculada. Matematicamente,

$$e(t) = d(t) - y(t). \quad (9)$$

Os pesos da próxima iteração, $t + 1$, são ajustados de acordo com a equação

$$w_m(t + 1) = w_m(t) + \eta e(t)x_m(t), \quad (10)$$

onde η é a taxa de aprendizado, até que o valor do erro esteja dentro de uma margem de tolerância definida pelo projetista. Um exemplo típico deste tipo de ANN são as redes Perceptron Multi-Camadas ou MLP (acrônimo do inglês *Multi-Layer Perceptrons*) (20).

Por outro lado, há casos nos quais somente os dados de entrada estão disponíveis para a rede. Nesta situação, a MLP é projetada para coletar estatísticas relevantes dos padrões da entrada de modo a agrupar os dados em classes pré-definidas. Dizemos, portanto, que a rede possui *aprendizado não-supervisionado* (21). As Redes de Kohonen, também conhecidas como Mapas Auto-Organizáveis ou SOMs (acrônimo do inglês *Self-Organizing Maps*), são um exemplo de ANN com aprendizado não-supervisionado (17).

2.3 Redes Neurais Convolucionais

Uma *Rede Neural Convolucional* ou CNN (acrônimo do inglês *Convolutional Neural Network*) ou, ainda, ConvNet é um tipo de ANN do tipo *feed-forward*⁷, cuja arquitetura é dotada de uma variação da MLP projetada para um custo computacional o mínimo (22) e inspirada no funcionamento do córtex visual humano (23).

Essas redes também são conhecidas como ANNs Invariantes a Deslocamento, ou SIANN (acrônimo do inglês *Shift-Invariant Artificial Neural Network*), pois o resultado do processamento independe da posição espacial dos objetos na imagem (24). Desta forma, possuem performance considerada estado-da-arte em processamento de imagens estáticas e sequenciais (vídeos).

⁷ Em ANNs do tipo *feed-forward*, a informação de saída de um neurônio só é utilizada pelos neurônios da camada posterior (21).

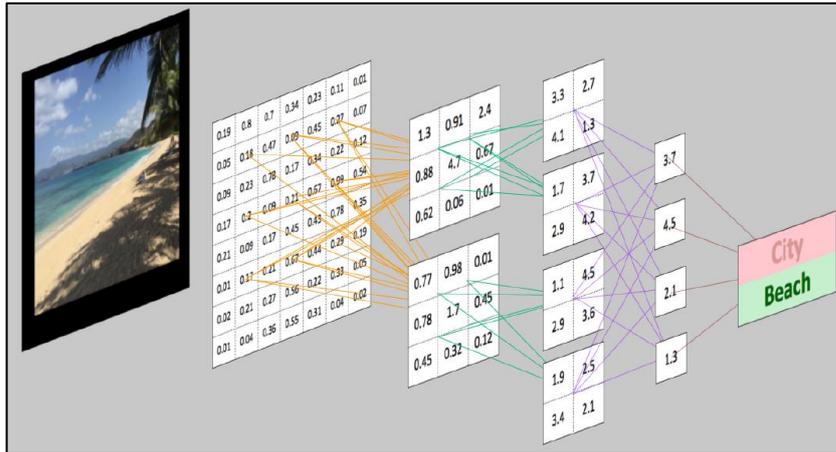


Figura 8. Ilustração resumida de uma CNN classificando uma imagem como “cidade” ou “praia”.

Apesar de ser aplicada em diversas áreas, como processamento de voz e linguagem natural (25), as CNNs são amplamente utilizadas no processamento de imagens e vídeo. Podemos citar como trabalhos relevantes o de Ferreira (26), que utilizou CNNs para detecção de plantas daninhas em imagens de *drone* em plantações de soja; ou o de Pereira et al (27), que desenvolveram uma arquitetura de CNN capaz de segmentar tumores cerebrais em imagens de ressonância magnética com escores suficientes para vencer o *Brain Tumor Segmentation Challenge 2013*.

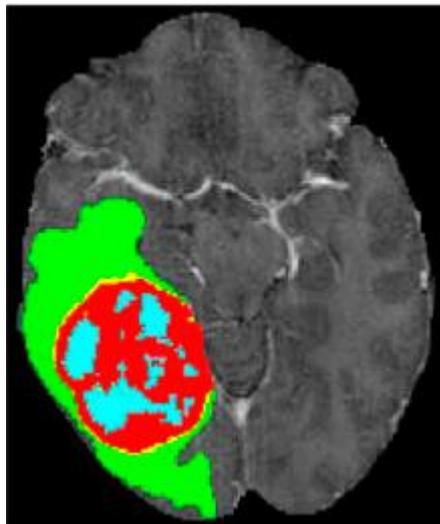


Figura 9. Imagem de ressonância magnética de um tumor cerebral segmentado e classificado por uma CNN, dividido por classes: tumor em crescimento (vermelho), tumor de tamanho estabilizado (amarelo), azul (necrose), verde (edema) (27).

Nas seções a seguir, serão apresentados os conceitos mais importantes para a compreensão dos elementos e do funcionamento das CNNs, bem como as arquiteturas utilizadas neste estudo.

2.3.1 HIPERPARÂMETROS

Hiperparâmetros são configurações escolhidas pelo projetista da CNN e que juntas determinam a performance da mesma (22).

Na seção 2.1.3, foram apresentados os conceitos de alguns dos principais hiperparâmetros, como filtro e *zero-padding*. Conforme definem Patterson e Gibson (22), os demais de maior relevância serão brevemente descritos a seguir.

2.3.1.1 Passo

O *passo* ou *stride* (representado aqui matematicamente por S) é um hiperparâmetro que especifica quantos *pixels* por vez o filtro se move à medida que convolui. No exemplo da Figura 6, o passo adotado foi de 1.

A escolha do passo utilizado dependerá da aplicação prática do resultado obtido. O projetista deverá ter em mente que passos maiores produzirão volumes de saída menores espacialmente.

2.3.1.2 Profundidade

A *profundidade* ou *depth* corresponde ao número de filtros utilizados, onde cada um deles se especializa em analisar aspectos diferentes da imagem.

Se a primeira camada convolucional (definida adiante) tomar como entrada uma imagem bruta de dimensões $W \times H$, o volume de saída terá profundidade D correspondendo à sobreposição de D filtros com habilidades de processamento individuais (detecção de bordas, análise de cor, etc.). A coluna formada neurônios espacialmente correspondentes ao longo da

profundidade D enxergam o mesmo *campo de visão*⁸ e juntos formam a *coluna de profundidade*.

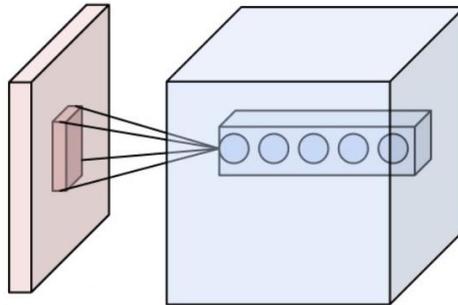


Figura 10. Para cada área da imagem, a convolução do filtro alimenta um neurônio por camada do volume de saída. Neste caso, é representada uma coluna de profundidade de 5 neurônios do volume de saída, na qual cada neurônio é resultado da convolução do filtro associado sobre uma área específica.

2.3.1.3 Tamanho da camada

O número de neurônios em uma determinada camada determina seu *tamanho*. A quantidade de neurônios no início e no fim da CNN não costuma ser um desafio, pois normalmente correspondem respectivamente ao número de entradas e saídas do problema.

A complexidade do projeto de CNNs reside nas camadas ocultas, uma vez que não há nenhuma regra quantificando essas camadas de acordo com o problema. O projetista deve encontrar a quantidade adequada de neurônios para formar uma rede eficaz, sempre tendo em mente que cada quantidade tem um custo computacional.

2.3.1.4 Taxa de aprendizagem

A *taxa de aprendizado* é uma constante de proporcionalidade que determina o ritmo de decremento da função de erro global em direção ao valor máximo de erro tolerável.

Taxas de aprendizado muito altas podem implicar em um avanço brusco em direção ao objetivo, causando risco de descarte de uma solução ótima para o problema ou de um

⁸ O *campo de visão* é a área processada por um passo de uma convolução, cujo resultado é armazenado em um neurônio (24).

treinamento instável e que não converge ao longo do tempo.

Por outro lado, uma taxa de aprendizado muito pequena poderá levar demasiado tempo para conclusão do processo de treinamento e tornar o algoritmo de aprendizado ineficiente.

As taxas de aprendizado representam, também, certo desafio para o projetista pois acabam sendo específicas para o conjunto de dados (e até para outros hiperparâmetros), tornando a obtenção de uma taxa de aprendizado adequada uma tarefa eventualmente cansativa.

2.3.1.5 Função de ativação

Servem para determinar matematicamente se um neurônio deve ou não ser ativado em função da relevância da informação transmitida a ele naquele momento.

Em outras palavras, utilizam-se *funções de ativação* em neurônios ocultos em uma rede neural para introduzir a não-linearidade nos recursos de modelagem da rede. Uma ANN desprovida de função de ativação é limitada a resolver apenas problemas de solução linear.

Nas CNNs, a função de ativação mais utilizada é a ReLU (acrônimo do inglês *Rectified Linear Unit*), dada por

$$\varphi(x) = x^+ = \max(0, x), \quad (11)$$

cuja discussão pode ser encontrada na obra de Googfellow et al (15) e no trabalho de Ramachandran et al (28).

2.3.2 CAMADA CONVOLUCIONAL

A *camada convolucional* é responsável por extrair características do volume de entrada através do uso de filtros que são convoluídos sobre o volume, gerando *mapas de características*. Esta camada da CNN é responsável pela parte mais pesada do trabalho computacional.

O mapa de características é uma estrutura bidimensional que acomoda as respostas desse filtro em todas as posições espaciais da imagem de entrada da camada. A CNN aprenderá

que filtros são ativados quando detectam alguma característica específica, como bordas, manchas, ou eventualmente padrões completos.

Tem-se um conjunto inteiro de filtros em cada camada convolucional, e cada um deles produzirá um mapa de características bidimensional separado, que são empilhados ao longo da dimensão de profundidade, produzindo o volume de saída. Esta ideia é ilustrada na Figura 11.

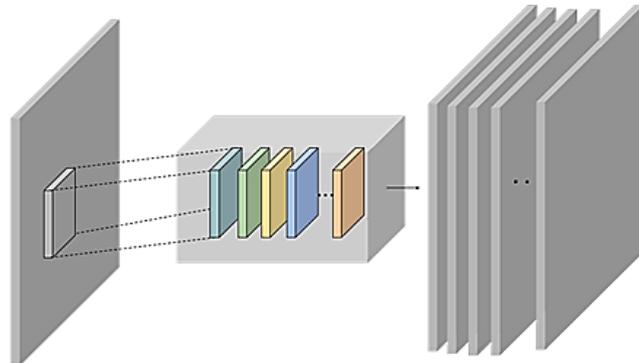


Figura 11. A convolução dos N filtros sobre a imagem gera N camadas resultantes, que juntas formam o volume de saída.

2.3.3 CAMADA DE POOLING

Em uma arquitetura CNN, é comum a inserção periódica de camadas de *pooling* entre as sucessivas camadas de convolução. Sua função é reduzir progressivamente o tamanho espacial da representação de modo a reduzir a quantidade de parâmetros a serem processados, mantendo, portanto, o controle do *overfitting*⁹.

Assim como na camada de convolução, a camada de *pooling* é obtida através da aplicação de uma máscara sobre a imagem. A função da máscara pode variar de acordo com a necessidade prática da CNN, podendo ser do tipo *MAX*, *AVERAGE*, entre outras.

A camada de *pooling* opera independentemente em cada canal da entrada e o redimensiona espacialmente, exceto a profundidade. De modo específico, a camada de *pooling* recebe um volume de tamanho $W_0 \times H_0 \times D_0$, requer o tamanho da máscara F e o passo S como hiperparâmetros, e com isto produz um novo volume de tamanho $W \times H \times D$, onde

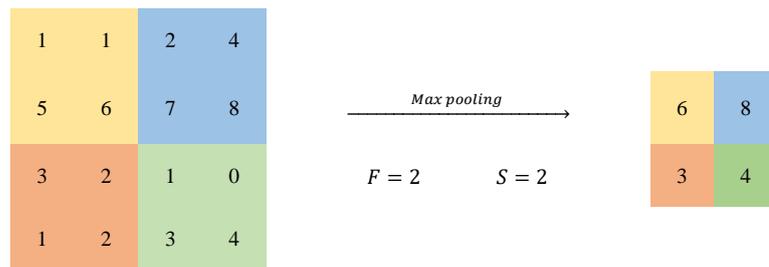
⁹ *Overfitting*, também chamado de sobreajuste, é um conceito estatístico que descreve o excelente ajuste de um modelo a um conjunto de dados anteriormente, mas se mostra ineficaz para prever novos resultados (52).

$$W = \frac{W_0 - F}{S + 1}, \quad (12)$$

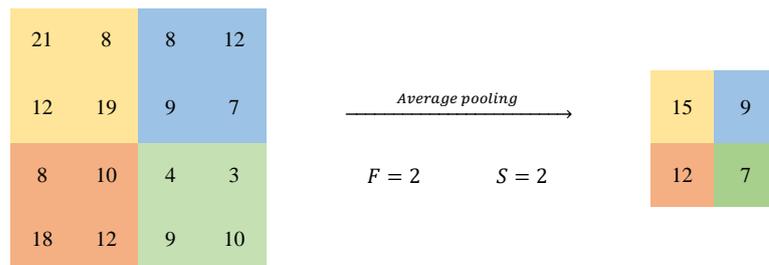
$$H = \frac{H_0 - F}{S + 1} e \quad (13)$$

$$D = D_0. \quad (14)$$

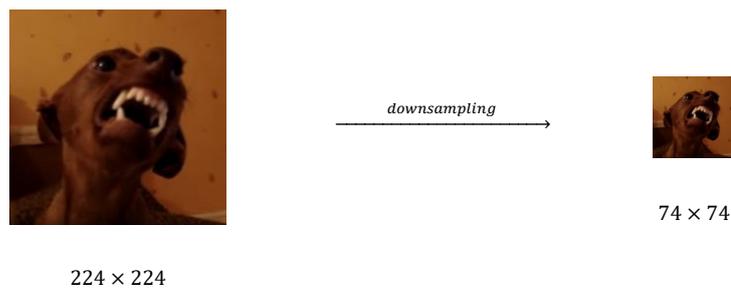
A Figura 12 exemplifica o processo de *pooling*.



(a)



(b)



(c)

Figura 12. a) *Pooling* através da função MAX, que preserva o maior valor dentro da máscara. b) *Pooling* através da função AVERAGE, que mantém a média dos valores interiores à máscara. c) Exemplo de *downsampling* de uma imagem RGB após *max pooling* da imagem. A quantidade de canais ($D = 3$) foi preservada.

2.3.4 ARQUITETURAS PARA SEGMENTAÇÃO DE IMAGENS

Em PDI, as CNNs são amplamente usadas em classificação (4) (19) (29) (30) (31) e segmentação (26) (27) (32) (33) (34) (35) (36) de imagens. Nas CNNs para classificação de imagens é comum ter uma ou mais *camadas totalmente conectadas*¹⁰ após as sucessivas camadas de convolução e de *pooling* (se for o caso), na qual a última camada possui número de neurônios correspondente à quantidade de classes da imagem. Por outro lado, as CNNs projetadas para segmentação de imagens não possuem camadas totalmente conectadas e são referidas como *Redes Totalmente Convolucionais* ou FCNs (acrônimo do inglês *Fully Convolutional Networks*).

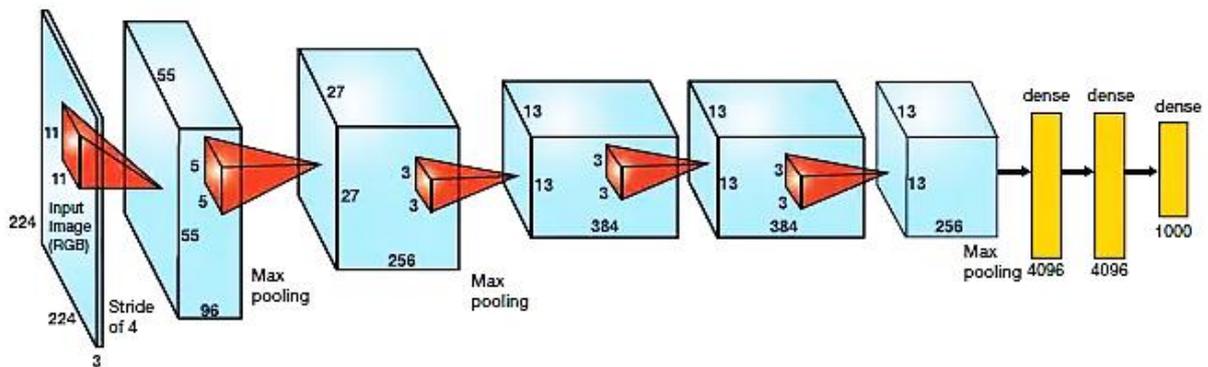


Figura 13. Arquitetura da AlexNet, uma CNN comumente usada para classificação de imagens (31).

O sucesso das CNNs na classificação de imagens motivou seu uso em problemas mais complexos, como a segmentação semântica de objetos. Diferentemente da classificação de imagens, a segmentação semântica é um problema de previsão estruturada onde cada *pixel* na grade da imagem precisa ser atribuído a um rótulo da classe à qual pertence (vias, telhado, água, terra, dentre outros exemplos).

O trabalho de Long et al (37) introduziu a ideia básica de espelhamento *downsampling-upsampling* das camadas de convolução, de modo que as previsões tivessem uma correspondência um-para-um com a imagem de entrada tendo em vista sua dimensão espacial ($W \times H$). Em outras palavras, dada uma posição na imagem, o resultado correspondente na mesma posição na grade de saída será uma previsão de categoria do pixel associado. Esta ideia

¹⁰ *Camadas totalmente conectadas* ou *camadas densas* são aquelas nas quais todo neurônio está conectado com todos os demais da próxima camada, seguindo a mesma definição das redes Perceptron Multicamada (17) (20).

é ilustrada na Figura 14.

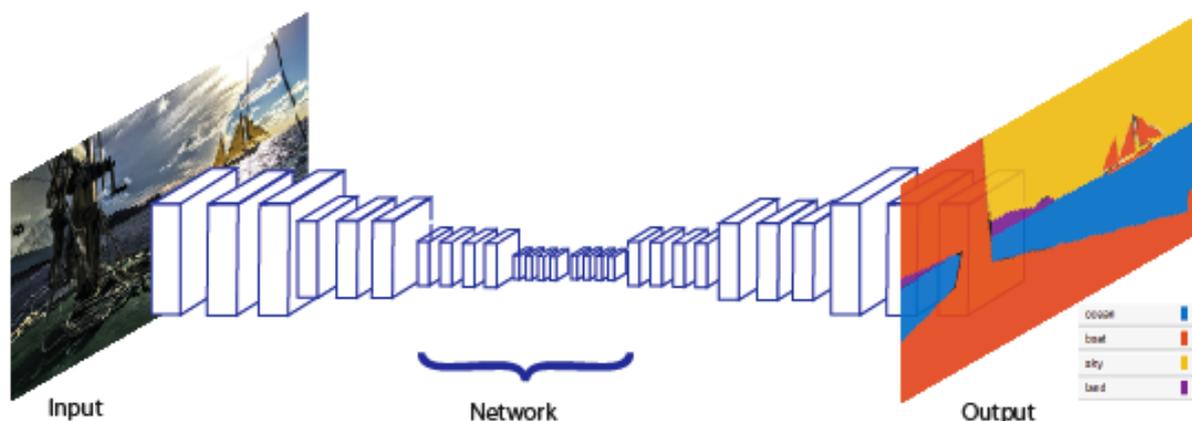


Figura 14. Esquema geral de uma segmentação semântica (38).

As arquiteturas utilizadas neste experimento são do tipo FCN e serão descritas nas seções a seguir.

2.3.4.1 U-Net

Conforme ilustrou LeCun et al (39), as primeiras CNNs requeriam enormes bases de dados para que os resultados obtidos estivessem dentro dos parâmetros esperados de tolerância. Objetivando melhorar este aspecto, a arquitetura *U-Net* foi desenvolvida por Ronneberger et al para obter resultados similares com bases de dados significativamente menores (32).

A parte *downsampling* (ou *encoding*) da U-Net, é similar a uma arquitetura típica de uma rede convolucional, porém sem as camadas densas. A ideia básica é posteriormente acrescentar uma rede de expansão, simétrica à de contração na qual as operações de *pooling* (seta vermelha na Figura 15) são substituídas por operadores *upsampling*, aumentando a resolução da saída até que seja igualada à de entrada.

Esta etapa de *upsampling* (ou *decoding*) consiste basicamente em dois artifícios. O primeiro são operações de concatenação da imagem decodificada com a correspondente na etapa de *downsampling*, produzindo uma imagem contendo várias informações de contexto (mapa de características) que ainda não foram perdidas na etapa anterior (seta cinza na Figura

15). O segundo são operações regulares de *convolução transposta*¹¹, que ocorrem até que a imagem resultante possua as mesmas dimensões da imagem de entrada.

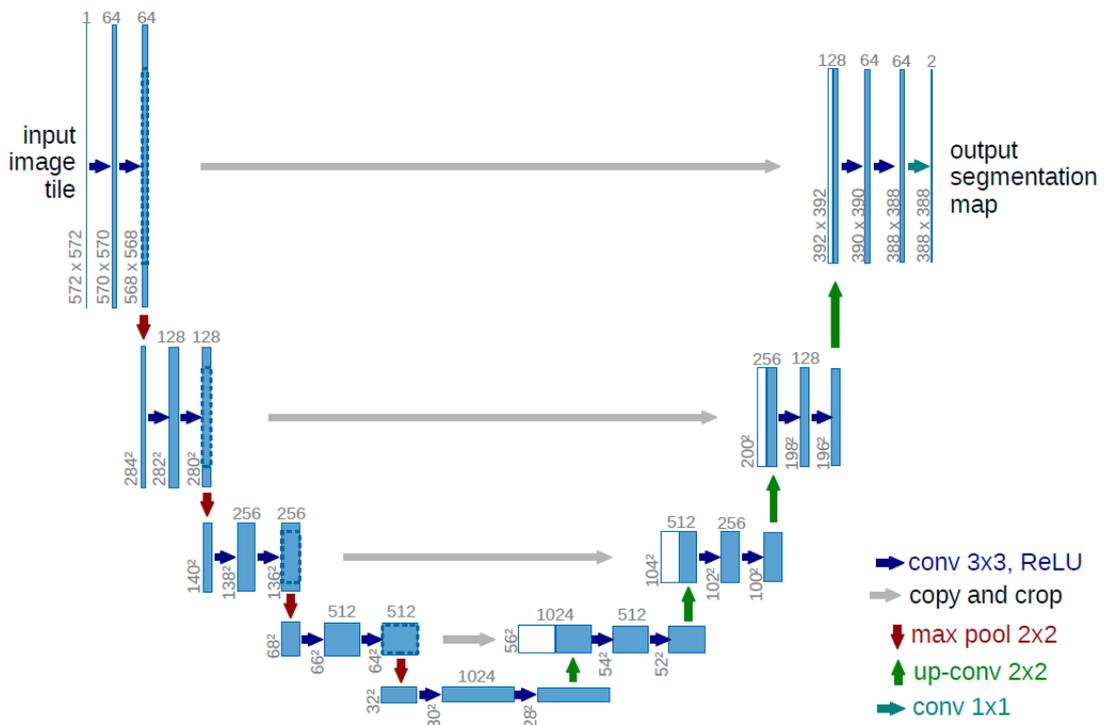


Figura 15. Arquitetura U-Net. Cada caixa azul corresponde a um mapa de características multicanal. O número de canais é indicado na parte superior da caixa. As dimensões da imagem ao longo do processo são indicadas no canto inferior esquerdo da caixa. Caixas brancas representam mapas de características copiados. As setas denotam as diferentes operações ao longo do processo (32).

A arquitetura proposta pelos autores obteve *IoU* médio de 92% contra 83% do segundo melhor algoritmo, quando testados no conjunto de dados “PhC-U373”, e de 77,5% contra 46% do sendo melhor algoritmo, quando testados no conjunto de dados “DIC-HeLa” (32).

2.3.4.2 Segnet

Baseando-se, também, no trabalho de Long et al (37), Badrinarayanan et al (33) desenvolveram a CNN *SegNet*. Da mesma forma que a CNN que a inspirou, a SegNet é

¹¹ A *deconvolução transposta* é caracterizada pelo uso da convolução em operações de *upsampling*, usualmente feitas a partir da aplicação de um filtro de tamanho $K = 3$ e passo $S = 2$ numa imagem com *zero-padding* $p = 1$. Cada elemento da imagem de origem está no centro do filtro em um passo associado da convolução (37).

constituída por duas partes: o *encoder*, similar à VGG-16 (40), na qual as operações de convolução e *pooling* são feitas, e o *decoder*, similar ao da U-Net.

A grande diferença entre U-Net e SegNet é que esta, durante a etapa de *downsampling*, armazena os índices dos *pixels* selecionados pelo *max pooling*. Então, através da convolução transposta, os *pixels* armazenados (em vez do mapa de características inteiro para concatenação) formam a imagem resultante. Isto proporciona uma maior preservação das características da imagem durante o processo de codificação e decodificação e uma economia substancial de memória e armazenamento.

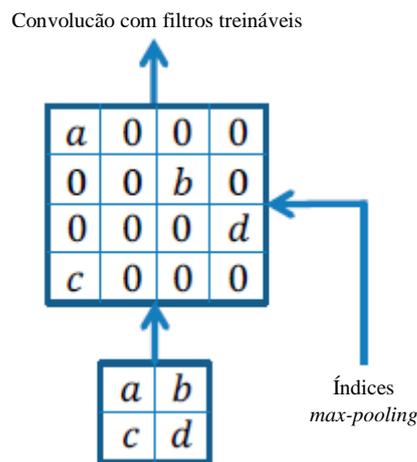


Figura 16. Representação do decodificador SegNet. Os índices são utilizados para montar o próprio mapa de características a ser convolvido por um filtro decodificador treinável (33).

Esta CNN obteve *IoU* médio de 60,1% quando testados no conjunto de dados “CamVid” concorrentes não tiveram esta métrica extraída) e 90,4% de acurácia contra 83,8% do segundo melhor algoritmo nesta mesma base de dados. Já na base “SUN RGB-D”, o *IoU* médio foi de 31,84% contra 32,08% do melhor algoritmo, porém superando todos em acurácia (33).

2.3.4.3 Dilated Convolutions

As arquiteturas percorridas no trabalho de Ronneberger et al (32) e no de Badrinarayanan et al (33), ou mesmo as apresentadas em outros estudos (7) (37) eram CNNs

originalmente criadas para classificação de imagens e adaptadas para segmentação de imagens. Por outro lado, Yu e Koltun (34) apresentaram em seu trabalho uma arquitetura de CNN específica para este propósito, baseada em convoluções dilatadas e agregação de contexto multi-escala.

A convolução β -dilatada é uma generalização da equação (6) para um fator de dilatação β qualquer, e é definida como

$$g_{\beta}(x, y) = f(x, y) *_{\beta} h(x, y) = \sum_{u=-F_1}^{F_1} \sum_{v=-F_2}^{F_2} f(u, v) \cdot h(x - \beta u, y - \beta v). \quad (15)$$

onde $\beta \in \mathbb{N}$.

A implementação desta operação proporciona a possibilidade de uma expansão exponencial do campo receptivo, conforme ilustra a Figura 17, e dada matematicamente por

$$F_{u+1, v+1} = (2^{u+2} - 1, 2^{v+2} - 1). \quad (16)$$

Como os filtros usualmente são quadrados, tem-se nestes casos que $u = v = k$. Podemos dizer, portanto, que o filtro F_{k+1} possui dimensões $(2^{k+1} - 1) \times (2^{k+1} - 1)$. O processo de convolução ocorre tal como descrito na seção 2.1.3, com as devidas adaptações.

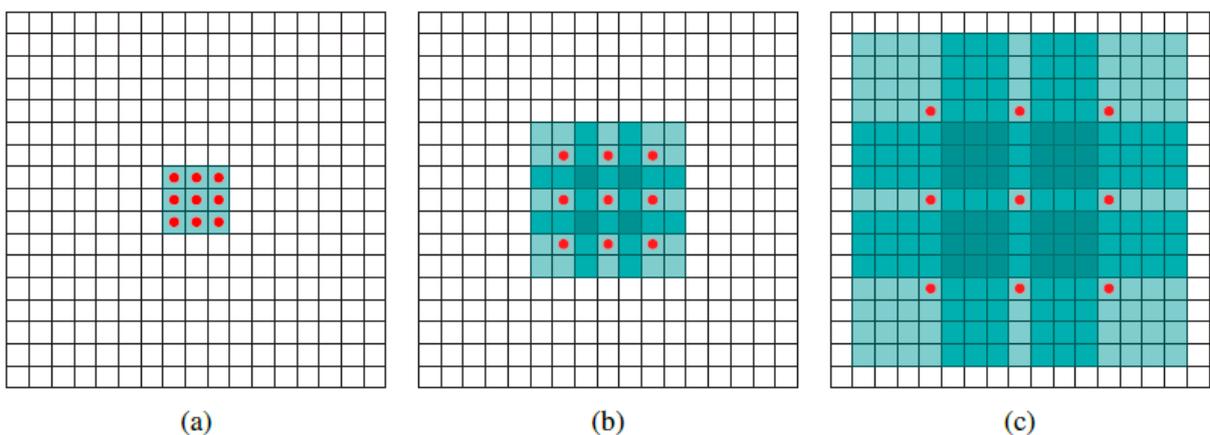


Figura 17. A dilatação suporta uma expansão exponencial do campo receptivo sem perda de resolução. a) Filtro para convolução 1-dilatada ($k = 3$) (a), 2-dilatada ($k = 7$) (b) e 3-dilatada ($k = 15$) (c) (34).

Tendo por base a convolução dilatada, o módulo de contexto é composto por 7 camadas que aplicam convoluções 3×3 com diferentes fatores de dilatação, a saber: 1, 1, 2, 4,

8, 16 e 1. Adicionalmente, há uma última camada aplicando uma convolução 1×1 , que mapeia o volume de saída da camada 7 para o formato da imagem de entrada. Este esquema é resumido na Tabela 1.

Tabela 1. Arquitetura da rede de contexto. A rede processa C mapas de características agregando informação contextual em escalas progressivamente crescentes sem perder a resolução (34).

Camada	1	2	3	4	5	6	7	8
Fator de dilatação (β)	1	1	2	4	8	16	1	1
Truncation	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Não
Tamanho do campo receptivo	3×3	5×5	9×9	17×17	33×33	65×65	67×67	67×67

A arquitetura proposta pelos autores obteve um IoU médio de 67,6% quando testada na base de dados “PASCAL VOC 2012”, superando outras arquiteturas já existentes, como FCN-8s (62,2) (37) e DeepLab (62,9) (35).

Capítulo 3

Materiais e métodos

3.1 Visão geral

O estudo em vista consiste em cumprir uma série de etapas que sequencialmente culminam no resultado esperado e é ilustrado na Figura 18.

A primeira etapa consiste na obtenção das imagens para treino e teste das CNNs. Em seguida, na segunda etapa, definem-se as regiões de interesse e os conjuntos de imagens de treinamento e teste. Já a terceira etapa consiste em treinar as CNNs para geração de suas bases de conhecimento, tornando-as capazes de executar a tarefa proposta. Finalmente, na quarta etapa, ao aplicar CNNs treinadas na etapa anterior, obtêm-se a segmentação das vias nas imagens fornecidas.

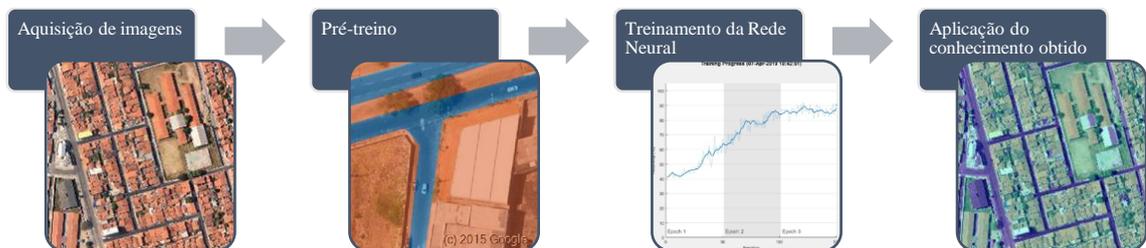


Figura 18. Etapas da pesquisa proposta em sequência.

As seções a seguir detalham as especificidades e procedimentos de cada etapa da sequência apresentada.

3.1.1 AQUISIÇÃO DE IMAGENS

Para aplicações automatizadas a banda térmica da imagem é um importante diferencial, uma vez que a reflectância¹² é uma variável altamente relevante na diferenciação¹³ e classificação de objetos em imagens de sensoriamento remoto (Figura 19). Entretanto, em virtude da conhecida dificuldade na aquisição de imagens multiespectrais de alta resolução espacial, é comum que grande parte dos trabalhos de edição vetorial sejam feitos manualmente, com base em imagens de geoserviços gratuitos (como Google e Bing) e, portanto, sem a banda térmica.

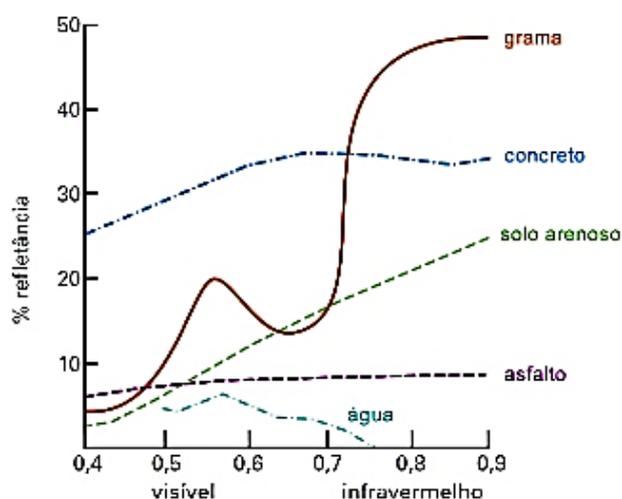


Figura 19. Percentual de reflectância das classes mais comuns em paisagens urbanas (41). É nítida sua diferenciação no espectro infravermelho, não presente em grande parte dos geoserviços gratuitos.

Tendo isto em vista, foram utilizadas 600 imagens RGB, de dimensões 256×256 pixels e resolução espacial de 25 cm, de modo que facilitasse ao máximo a detecção das vias e que houvesse uma quantidade mínima de pixels mistos, isto é, aqueles que pertencem a diversas classes na imagem. As imagens são de setores urbanos de São Luís-MA ($2^\circ 31' 48''$ S, $44^\circ 18' 10''$ O (42)), foram capturadas pela DigitalGlobe e publicadas pela Google.

¹² A reflectância é definida como a razão entre a energia no comprimento de onda refletida pelo objeto e a incidente sobre o mesmo (50).

¹³ Conforme demonstra Nóbrega em seu trabalho (4), a reflectância é uma das propriedades que proporciona a evidência de descontinuidades entre os diferentes objetos na imagem.

3.2 Base de Dados e pré-treino

Neste método, o treinamento da CNN é supervisionado, isto é, acontece a partir de uma base de dados previamente fornecida (43). Desta forma, deve ser antecedido por algumas etapas cuja finalidade é subsidiar e organizar informações para que a rede seja capaz de gerar uma base de conhecimento.

Para cada imagem é definida uma ROI, representada por uma matriz de inteiros, de mesmas dimensões. Cada célula (x, y) desta matriz está associada a um *pixel* da imagem e guarda um número natural que está associado a uma classe C , dada por

$$C(x, y) = \begin{cases} 0, & \text{se o } \textit{pixel} \text{ não pertence a nenhuma classe;} \\ 1, & \text{se o } \textit{pixel} \text{ corresponde a uma via;} \\ 2, & \text{se o } \textit{pixel} \text{ pertence a outra classe que não seja uma via.} \end{cases}$$

As ROIs foram desenhadas com o auxílio do *Matlab Image Labeler*.

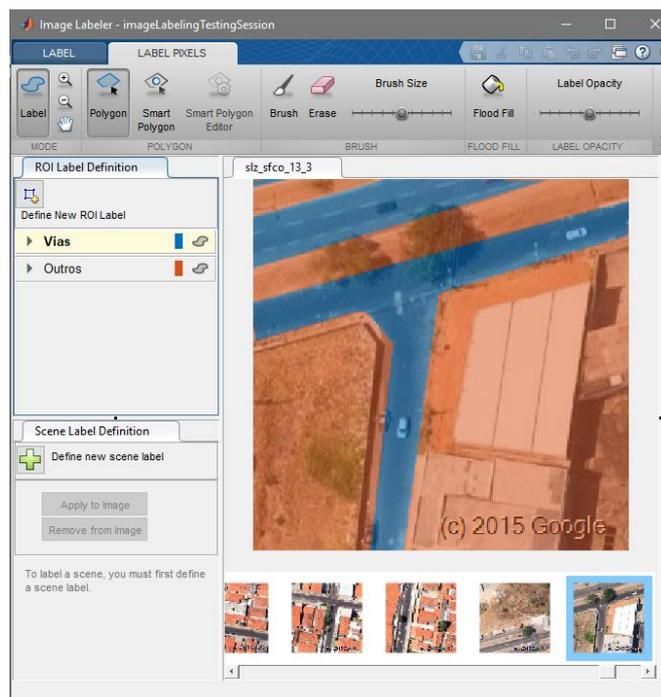


Figura 20. Edição de uma ROI no *Matlab Image Labeler*. Na figura ao centro, os *pixels* correspondentes à classe 1 (vias) estão em azul e os à classe 2 (outros) estão em laranja.

Contudo, é importante ressaltar que, por uma questão de praticidade e viabilidade, as ROIs foram traçadas sem levar em conta alguns elementos poluidores da imagem, tais como quantidade considerável de sedimentos característicos¹⁴ do solo de São Luís, arborização aleatória sobre porções consideráveis de algumas vias, sombras de construções e arborizações, tons e texturas variados no asfalto ocasionados por pavimentação precária e remendos posteriores, a marca d'água gerada pelo próprio Google durante o *download* das imagens como condição para concessão das mesmas, e até mesmo diferenças de iluminação entre quadrantes obtidos em diferentes épocas pelo satélite.



Figura 21. a) Os diversos elementos poluidores da imagem destacados em retângulos: sedimentos arenosos sobre a pista (azul), arborização aleatória (vermelho), sombras (amarelo), tons de concreto em consertos de erosões no asfalto (verde) e marca d'água do Google (canto inferior direito). b) imagem do item a) com sua ROI desenhada sem considerar os referidos elementos poluidores.

Por fim, do total de imagens, 500 destas, com suas respectivas ROIs, serão destinadas ao treinamento da rede. As 100 restantes, nas mesmas condições, serão utilizadas pela rede para aplicação e validação da base de conhecimento gerada.

3.3 Treinamento da Rede Neural

O treinamento da rede neural consiste em gerar base de conhecimento a partir de uma

¹⁴ O solo da Grande São Luís é composto principalmente por latossolo e argissolo (48), o que torna natural a presença de sedimentos arenosos e argilosos na paisagem urbana em questão (49).

base de treino, isto é, calcular uma função-objetivo que, dada como entrada uma imagem como a especificada neste estudo, identifique quais *pixels* correspondem a uma via urbana.

O método de treinamento é supervisionado, ou seja, para cada entrada de dados, é fornecida a saída desejada. A rede utiliza essas informações para ajustar a função-objetivo de modo que, ao final do processo, a região de interesse seja corretamente determinada, dentro de parâmetros viáveis de precisão.

Os treinamentos da U-Net, SegNet e Dilated Convolutions foram feitos no *Matlab Deep Learning Toolbox*, utilizando uma máquina equipada com CPU Intel Core i7 7700HQ 3.6 GHz, GPU Nvidia GeForce 1060M GTX 6 GB GDDR5, SSD NVMe Samsung 960 Pro 1 TB e 2×SDRAM DDR4 Crucial 2400 MHz 16 GB, e outras configurações que são irrelevantes para o presente estudo.

3.4 Métricas de avaliação

Na avaliação deste estudo, considerou-se as seguintes variáveis (cuja relação entre pode ser ilustrada na Figura 22):

- VP: variável correspondente aos *Verdadeiros Positivos*, isto é, *pixels* detectados como correspondendo a uma via asfaltada, e que de fato pertencem a esta classe;
- FP: variável correspondente aos *Falsos Positivos*, isto é, *pixels* detectados como correspondendo a uma via asfaltada, mas que não pertencem a esta classe;
- VN: variável correspondente aos *Verdadeiros Negativos*, isto é, *pixels* detectados como não correspondente a uma via asfaltada, e que de fato não pertencem a esta classe;
- FN: variável correspondente aos *Falsos Negativos*, isto é, *pixels* detectados como não correspondente a uma via asfaltada, mas que na verdade pertencem a esta classe;

Conforme discorrido nos estudos de Csurka et al (44), estas variáveis servem de parâmetro para o cálculo de diversas métricas de qualidade, a saber:

- *Acurácia*: indica o percentual de *pixels* corretamente identificados. Esta

métrica indica o quão bem a CNN identifica os pixels correspondentes às vias asfaltadas. É dada por

$$Accuracy = \frac{VP}{VP + FN} \quad (17)$$

uma vez que o denominador da fórmula corresponde ao *ground truth*¹⁵. O valor de acurácia considerado será a média entre as acurácias obtidas para todas as imagens do conjunto de teste.

- *Índice de Jaccard*: também denominado Coeficiente de Similaridade de Jaccard, ou *IoU* (acrônimo do inglês *Intersection over Union*). Matematicamente, o *IoU* entre dois conjuntos *A* e *B* é dado por

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (18)$$

Esta métrica estatística considera e penaliza os falsos positivos. Escrevendo a equação (18) nos termos das variáveis definidas nesta seção, temos

$$IoU = \frac{VP}{VP + FP + FN} \quad (19)$$

O Índice de Jaccard considerado será a média entre os *IoUs* obtidos para todas as imagens do conjunto de teste.

- *Tempo de segmentação*: é o tempo gasto desde a chamada da função de segmentação até a finalização deste processo.

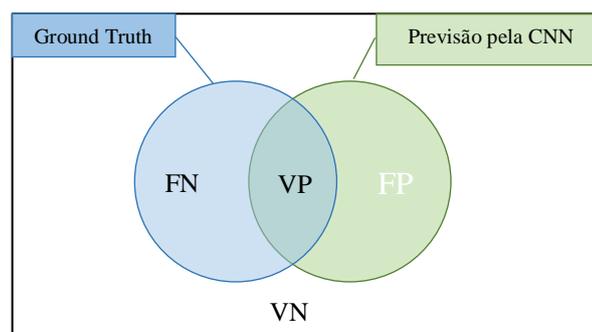


Figura 22. Diagrama de Venn para as variáveis preditivas.

¹⁵ *Ground truth* é o conjunto dos pixels que de fato correspondem a uma determinada classe na realidade (44).

Capítulo 4

Resultados e Discussões

4.1 Avaliação da U-Net

U-Net foi capaz de atingir uma acurácia de apenas 65,28% para a detecção de vias, porém alcançou uma precisão de 90,9% para o plano de fundo. Esta disparidade revela que esta CNN considerou boa parte dos *pixels* correspondentes às vias como não pertencentes a esta classe, gerando um quantitativo bastante expressivo de falsos negativos. Em contrapartida, houve baixo quantitativo de falsos negativos para o plano de fundo, mostrando que esta CNN identifica razoavelmente bem o que não faz parte das vias asfaltadas.

O pior caso obteve acurácia média de 67,24%, confirmando a dificuldade desta CNN em identificar falsos negativos como verdadeiros positivos. Além disso, obteve um *IoU* médio de 45,45%, que, quando confrontado com a acurácia, indica que o quantitativo de falsos negativos interfere na precisão da predição tanto quanto o quantitativo de falsos positivos. A presença relevante de falsos positivos indica uma certa dificuldade da U-Net em diferenciar pixels de diferentes objetos mas que possuem tons e textura similares, conforme consta na Figura 23-a.

Por outro lado, o melhor caso (Figura 23-b) obteve acurácia média de 92,13% e *IoU* médio de 86,46%, evidenciando o bom desempenho da rede quando não há tantos objetos com características similares ao asfalto, em especial cor e textura. É possível observar nas imagens originais correspondentes ao pior caso e ao caso mediano a presença de telhados de fibrocimento e lajes de concreto, que possuem cor e textura próximas às do asfalto. Já na imagem original correspondente ao melhor caso, há certa uniformidade nas características do

asfalto e, ao mesmo tempo, grande contraste entre as características dos demais objetos que compõem o plano de fundo.

O caso mediano da amostra (Figura 23-c) obteve 76,51% de acurácia média e 64,23% de *IoU* quando avaliado pela U-Net. Foram observadas as mesmas falhas do pior caso, embora em escala e frequência menores.

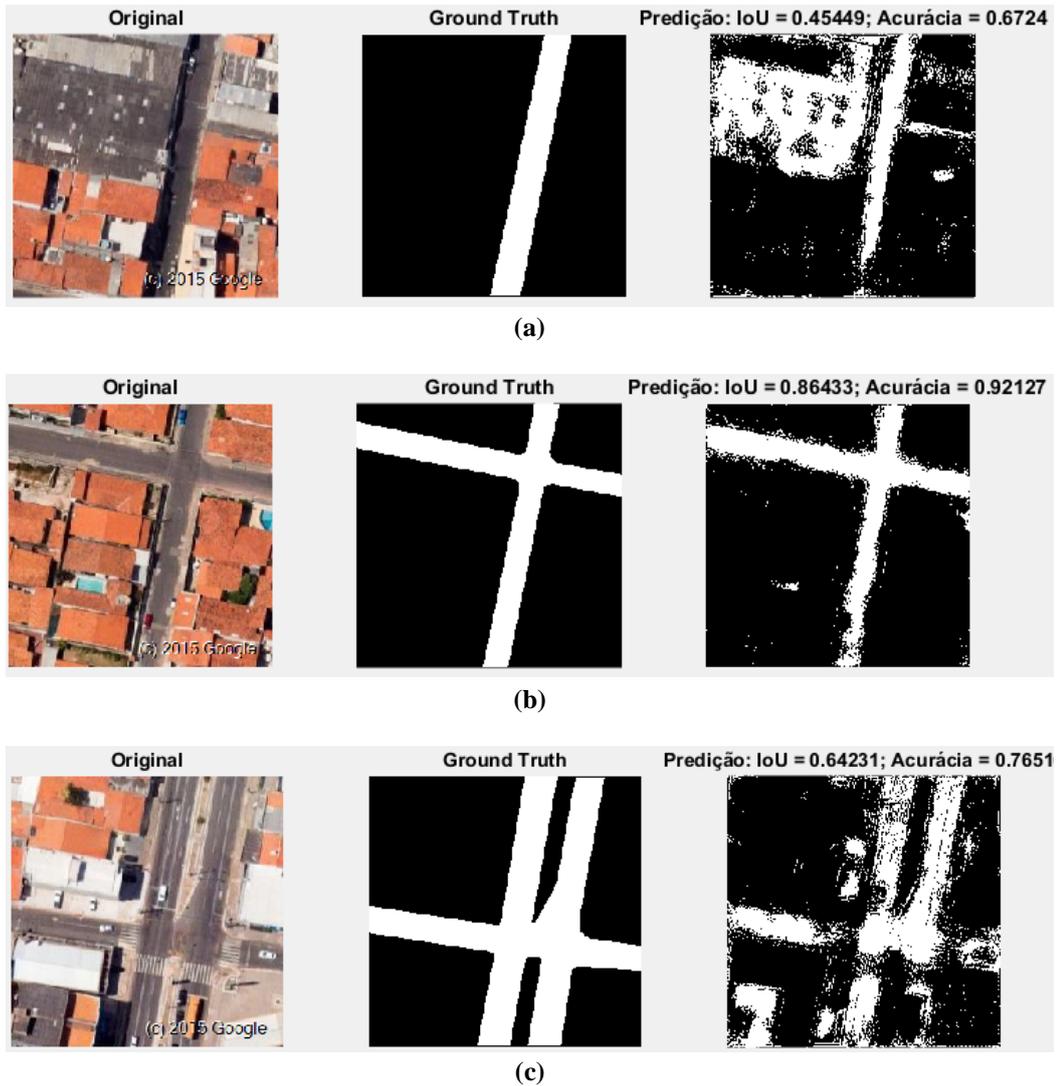


Figura 23. Pior (a), melhor (b) e mediano (c) caso de teste para a U-Net referente à classe “Vias”.

4.2 Avaliação da SegNet

A SegNet obteve uma acurácia de 83,97% para a detecção de vias e de 90,26% para o plano de fundo. Estes números demonstram que esta CNN foi capaz de fazer uma boa

diferenciação entre as vias asfaltadas e o restante da imagem na maioria dos casos de teste.

O pior caso alcançou acurácia de 73,57% e IoU de 44,96%, mostrando considerável quantidade de falsos positivos. Isto denota que esta CNN tem a mesma dificuldade da U-Net em contextualizar os objetos e, assim, diferenciar vias asfaltadas de elementos cuja cor e textura se assemelham bastante ao asfalto, conforme pode ser observado na Figura 24-a.

Já o melhor caso (Figura 24-b) obteve acurácia de 94,73% e IoU de 90,24%, evidenciando, assim como a U-Net, desempenho satisfatório da SegNet mesmo com alguma quantidade de elementos poluidores da imagem (neste caso, notadamente sombras e sedimentos). Ainda assim, nota-se um discreto quantitativo de falsos positivos no canto inferior direito da imagem, reforçando a denotação colocada na análise do pior caso.

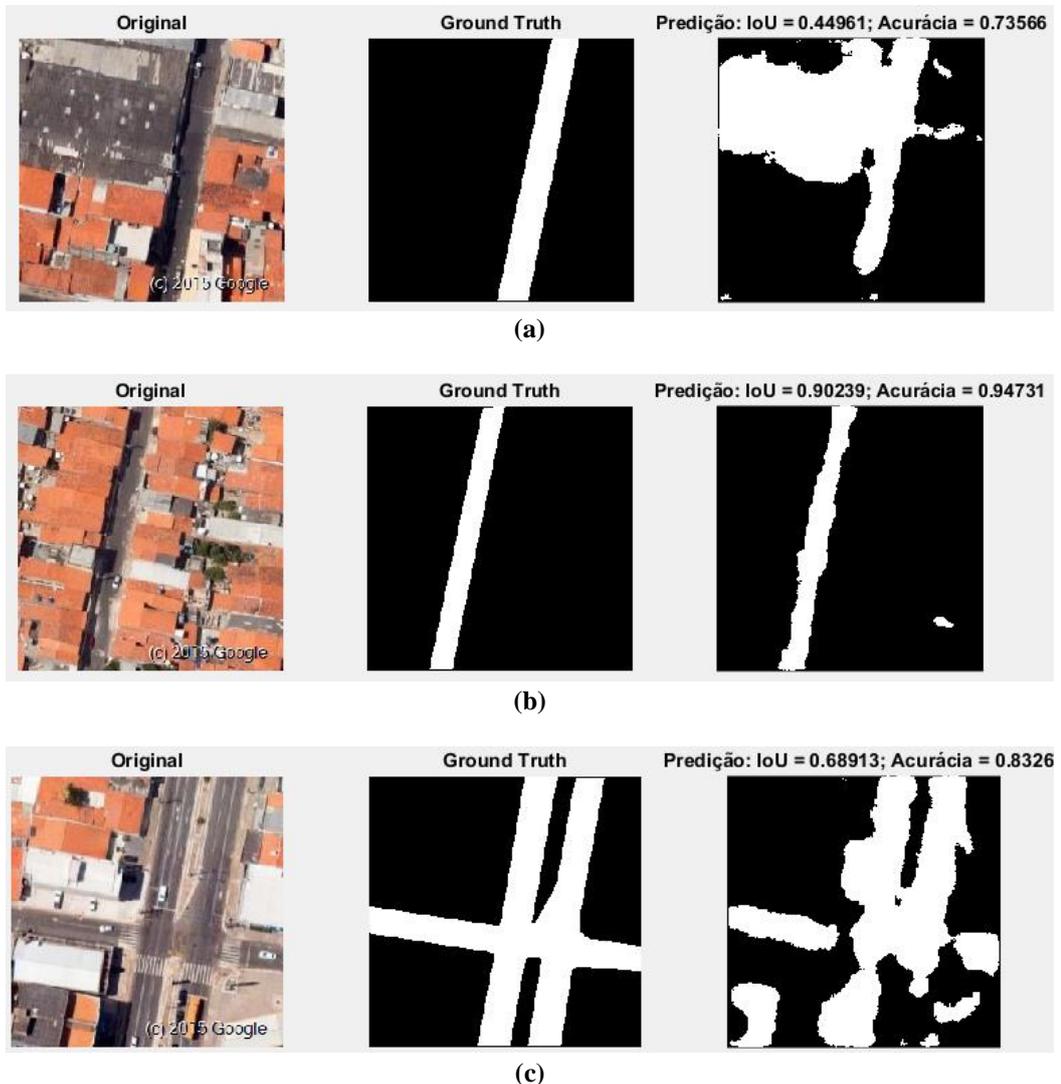


Figura 24. Pior (a), melhor (b) e mediano (c) caso de teste para a SegNet referente à classe “Vias”.

Similarmente, o caso mediano da amostra (Figura 24-c), aqui, obteve 83,26% de acurácia média e 68,91% de *IoU* quando avaliado pela SegNet.

4.3 Avaliação da Dilated Convolutions

A Dilated Convolutions atingiu uma acurácia de 91,36% para a detecção de vias e de 79,19% para o plano de fundo. Estes dados mostram que esta CNN também, assim como a SegNet, alcançou boa diferenciação entre as vias asfaltadas e o restante da imagem na maioria dos casos de teste. Contudo, o quantitativo de 20,81% de falsos positivos para o plano de fundo revela que a Dilated Convolutions apresenta certa dificuldade em delimitar o que não faz parte das vias, ocasionando o efeito oposto: geração de falsos negativos para a classe de vias, ocasionando baixa acurácia.

O pior caso teve uma acurácia de 67,96% e *IoU* de 29,76%, mostrando considerável quantidade de falsos positivos. Isto evidencia, também similarmente às CNNs anteriores, certa dificuldade desta CNN em considerar o contexto espacial os objetos, de modo a possibilitar a distinção das vias asfaltadas de elementos com cor e textura próximas aos do asfalto. Isto pode ser constatado na Figura 25-a.

Em compensação, o melhor caso, ilustrado na Figura 25-b, alcançou acurácia de 93,86% e *IoU* de 89,24%, mostrando desempenho próximo da SegNet em condições similares. Da mesma forma, mesmo no melhor caso é possível constatar discreto quantitativo de falsos positivos à esquerda e no canto superior direito da imagem, pelos motivos já discutidos.

Finalmente, a Figura 25-c corresponde ao caso mediano da amostra que, quando avaliado pela Dilated Convolutions, obteve 79,80% de acurácia média e 59,56% de *IoU*. Da mesma forma que na U-Net, é possível notar e comprovar a existência dos problemas já detectados no pior caso.

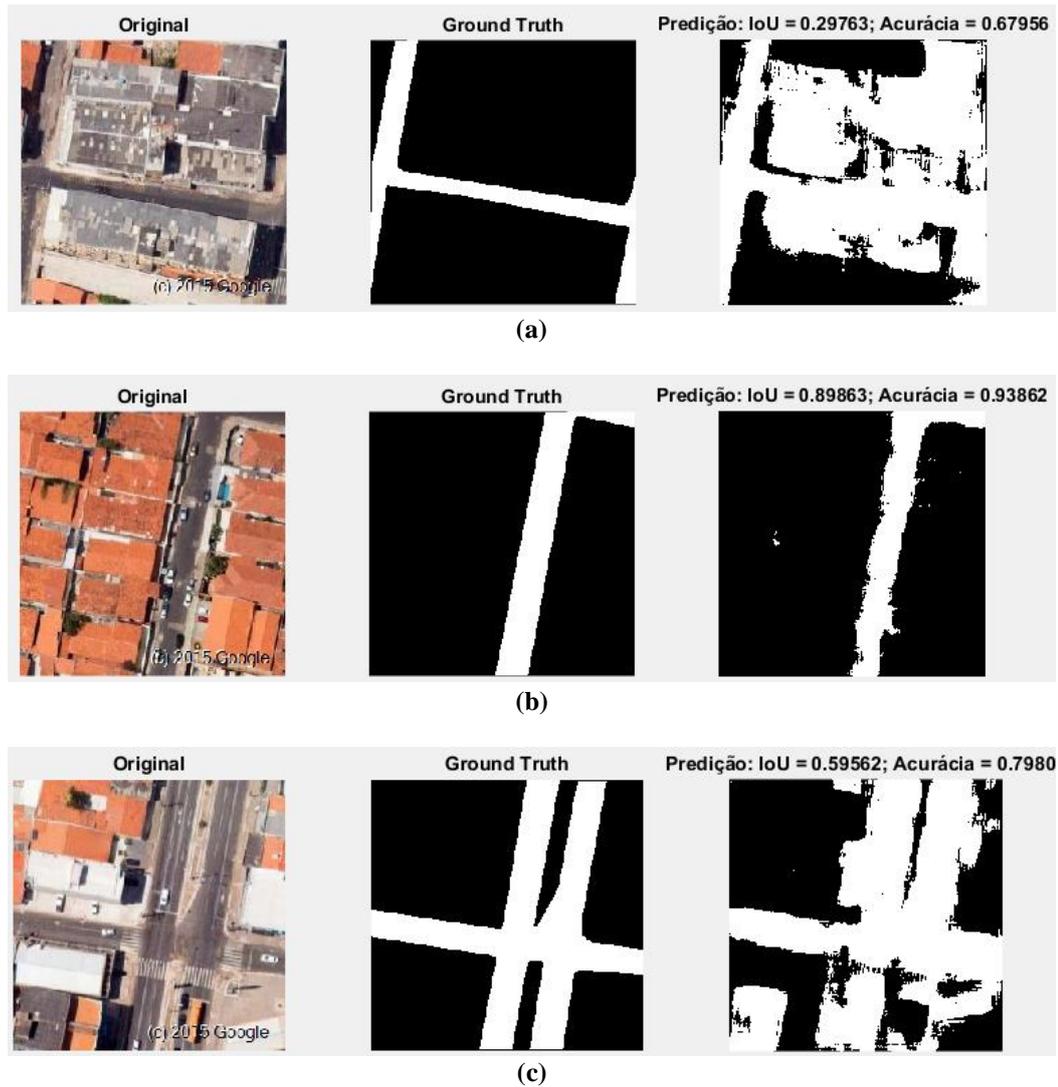


Figura 25. Pior (a), melhor (b) e mediano (c) caso de teste para a Dilated Convolutions referente à classe “Vias”.

4.4 Comparação entre U-Net, SegNet e Dilated Convolutions

Quando consideramos o contexto geral, isto é, levamos em consideração todas as classes em toda a base de treino, a U-Net alcança uma acurácia média de 78,09%, contra 87,12% da SegNet e 85,27% da Dilated Convolutions. Estes dados mostram que, nestas condições, a U-Net é menos adequada para a tarefa proposta que as outras duas em virtude da quantidade expressiva de falsos negativos. Na prática, em várias imagens parte notável das vias deixará de ser detectada como tal.

Incluindo os falsos positivos no conjunto, U-Net logra um *IoU* médio de 64,91%, enquanto SegNet marca 71,93% e Dilated Convolutions 61,27%. Apesar de aparentemente

estar na mesma média que a Dilated Convolutions no que se refere aos falsos positivos, a aplicação prática da U-Net fica comprometida em virtude da dificuldade desta CNN em avaliar corretamente os falsos negativos, provavelmente por ser mais sensível à natureza heterogênea do asfalto da área de estudo. Neste caso, SegNet mostra um melhor resultado prático em termos de ausência de classificações falsas.

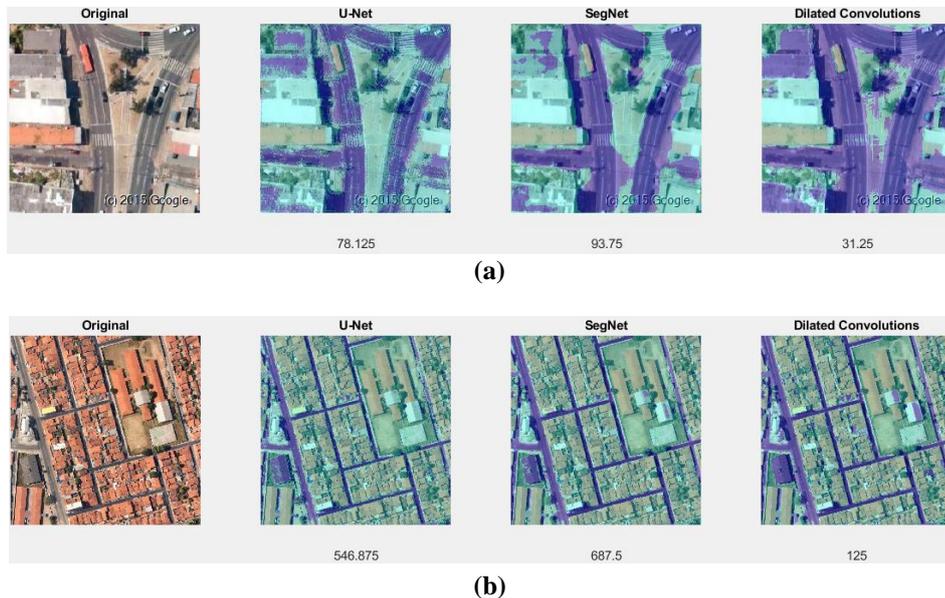


Figura 26. Comparativo do tempo médio (segundos) de segmentação de uma imagem aleatória da base de testes por cada CNN.

Tabela 2. Comparativo geral entre as CNNs U-Net, SegNet e Dilated Convolutions.

CNN	Acurácia (Vias/VP)	Acurácia (Outros/VN)	Acurácia Média	<i>IoU</i> Médio	Tempo médio (ms) (256 × 256 pixels)	Tempo médio (ms) (1024 × 1024 pixels)
U-Net	65,28%	90,90%	78,09%	64,91%	78,1	546,9
SegNet	83,97%	90,26%	87,12%	71,93%	93,7	687,5
Dilated Convolutions	91,36%	79,19%	85,27%	61,27%	31,2	125,0

Ao contrário das demais, a U-Net produz previsões com aspecto ruidoso. Isto sugere que as características mais ínfimas da imagem são preservadas durante o processo de *downsampling* e *upsampling*, embora na prática esta informação extra não se mostre útil. Por outro lado, apesar de marcar boa pontuação, SegNet apresenta grande facilidade em

descontinuar a região segmentada, o que pode gerar problema em certos tipos de aplicações, como as que requerem esqueletonizações¹⁶.

Quanto ao tempo de segmentação (Figura 26), U-Net completou a tarefa em 78,1 ms para uma imagem aleatória de 256×256 *pixels* e 546,9 ms para uma de 1024×1024 *pixels*. SegNet levou um tempo superior, de 93,7 ms, para a mesma imagem de 256×256 *pixels* e 687,5 ms para a mesma de 1024×1024 *pixels*. Por fim, para estas mesmas imagens, Dilated Convolutions completou a segmentação da imagem de 256×256 *pixels* em 31,2 ms e a de 1024×1024 *pixels* em 125 ms, demonstrando superioridade no tempo gasto para esta tarefa.

Os dados apresentados nesta seção podem ser sintetizados na

Tabela 2.

¹⁶ Uma discussão sobre o processo de esqueletonização pode ser encontrada no estudo de Plotze e Bruno (47).

Capítulo 5

Considerações Finais

Este trabalho contemplou em uma breve fundamentação teórica dos tópicos mais relevantes em Processamento Digital de Imagens para o entendimento do estudo apresentado. Também abordou tópicos relacionados às Redes Neurais Artificiais e Redes Neurais Convolucionais que serviram de base para o entendimento das três arquiteturas utilizadas neste experimento: U-Net, SegNet e Dilated Convolutions. A partir destas preliminares, apresentou-se a metodologia utilizada para execução do experimento e a discussão dos resultados obtidos.

Nas condições apresentadas, SegNet se mostrou mais apropriada a aplicações que exigem uma detecção precisa, pois obteve um desempenho superior na predição de vias asfaltadas com uma acurácia média de 87,12% e *IoU* médio de 71,93%, enquanto Dilated Convolutions alcançou os valores de 85,27% e 61,27% e U-Net 78,09% e 64,91%, respectivamente. Sua facilidade em descontinuar a área segmentada, porém, compromete sua indicação a aplicações sensíveis a este problema. Por outro lado, Dilated Convolutions confere maior adequação a aplicações que demandem um processamento de alta frequência e que não sofram tanta interferência dos falsos positivos, pois conseguiu balancear uma boa acurácia com um tempo médio bastante reduzido de 31,2 ms para imagens de 256×256 *pixels* e 125 ms para as de 1024×1024 *pixels*, contra os respectivos 78,1 e 546,9 da U-Net e 93,7 e 687,5 ms da SegNet. As CNNs analisadas neste experimento, nas circunstâncias apresentadas, demonstraram ter dificuldades em diferenciar objetos de cor e textura similares.

Trabalhos futuros podem estender este experimento a partir de sua melhoria em vários aspectos. As mais relevantes incluem: desenhar ROIs mais precisas, evitando que regiões não pertencentes às vias (em especial as citadas na seção 3.2) não agreguem conhecimento confuso ou impreciso ao treinamento; e aumentar o tamanho da base de dados, de forma que a CNN

disponha de maior variedade de informações para treino e, possivelmente, gerar conhecimento mais preciso. Um estudo mais aprofundado sobre a influência da configuração dos hiperparâmetros das CNNs no resultado da segmentação também pode fomentar melhores resultados. Vale comentar, ainda, que produções científicas mais recentes, como as de Guérin et al (45) e de Kanezaki (46), têm explorado o potencial das CNNs em métodos não-supervisionados de aprendizado e classificação, possibilitando, portanto, uma eventual comparação entre métodos supervisionados e não-supervisionados de segmentação de objetos baseados em redes convolucionais. Os métodos não-supervisionados não requerem *labels*, o que viabiliza o uso de bases de treino grandes já que não é necessário dedicar longo tempo e esforço no desenho de ROIs.

Além disso, o estudo apresentado tem potencial para diversas aplicações. Dentre elas, vale citar a vetorização automática da área segmentada utilizando algoritmos de esqueletonização, muitos deles discutidos e comparados no estudo de Plotze e Bruno (47). É relevante citar, também, a aplicabilidade em estudos que requeiram calcular a área correspondente às vias em paisagens urbanas. No que concerne à detecção de vias, também são encorajadas a investigação e a concepção de uma arquitetura de CNN específica para detecção de vias em imagens semelhantes às deste experimento. Finalmente, sugere-se avaliar o uso das CNN na detecção de vias de qualquer tipo, inclusive as de piçarra e terra, frequentes em áreas rurais e algumas áreas urbanas.

Referências

1. **Pinho, Carolina Moutinho Duque de, Feitosa, Flávia da Fonseca and Kux, Hermann J. H.** Classificação automática de cobertura do solo urbano em imagem IKONOS: Comparação entre a abordagem pixel-a-pixel e orientada a objetos. *Anais XII Simpósio Brasileiro de Sensoriamento Remoto*. abril 2005, pp. 4217-4224.
2. **Simões, Alexandre da Silva.** *Segmentação de imagens por classificação de cores: uma abordagem neural*. Escola Politécnica da Universidade de São Paulo. São Paulo : s.n., 2000. p. 171, Dissertação de Mestrado.
3. **Venturieri, Adriano.** *Segmentação de imagens e lógica nebulosa para treinamento de uma rede neural artificial na caracterização do uso da terra na região de Tucuruí (PA)*. Instituto Nacional de Pesquisas Espaciais. São José dos Campos : s.n., 1996. p. 140, Dissertação de Mestrado.
4. **Nóbrega, Rodrigo Affonso de Albuquerque.** *Detecção da malha viária na periferia urbana de São Paulo utilizando imagens de alta resolução espacial e classificação orientada a objetos*. Departamento de Engenharia de Transportes, Escola Politécnica da Universidade de São Paulo. São Paulo : EPUSP, 2007. p. 51, Tese de Doutorado.
5. **Pinho, Carolina Moutinho Duque de, et al.** Intra-urban land cover classification from high-resolution images using the C4.5 Algorithm. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. julho 2008, Vol. XXXVII, pp. 695-700.
6. **Doucette, Peter, et al.** Self-organised clustering for road extraction in classified imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2001, Vol. 55, 5-6, pp. 347-358.
7. **Noh, Hyeonwoo, Hong, Seunghoon and Han, Bohyung.** Learning Deconvolution Network for Semantic Segmentation. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. Dezembro 07, 2015, pp. 1520-1528.

8. **Gonzalez, Rafael C. and Woods, Richard E.** *Digital Image Processing*. 4ª. Upper Saddle River : Pearson, 2018.
9. **Reis, Artur Bernardo Silva.** *Metodologia computacional para a segmentação da próstata e classificação de lesões em imagens de ressonância magnética utilizando o Modelo de Ising*. Departamento de Informática, Universidade Federal do Maranhão. São Luís : s.n., 2019. p. 125, Tese de Doutorado.
10. **Braz Junior, Geraldo, et al.** Breast cancer detection in mammography using spatial diversity, geostatistics, and concave geometry. *Multimedia Tools and Applications*. Junho 25, 2018, pp. 1-27.
11. **Chatrath, Jatin, et al.** Real time human face detection and tracking. *2014 International Conference on Signal Processing and Integrated Networks*. Fevereiro 2014, pp. 705-710.
12. **George, Treesa, Potty, Sumi P. and Jose, Sneha.** Smile detection from still images using KNN algorithm. *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies*. Julho 2014, pp. 461-465.
13. **Sundararajan, D.** *Digital Image Processing: A Signal Processing and Algorithmic Approach*. Cingapura : Springer, 2017.
14. **Sonka, Milan, Hlavac, Vaclav and Boyle, Roger.** *Imaga Processing, Analysis, and Machine Vision*. 3ª. Toronto : Thomson Learning, 2008.
15. **Goodfellow, Ian, Bengio, Yoshua and Courville, Aaron.** *Deep Learning*. s.l. : MIT Press, 2016.
16. **dos Santos, Alexandre Rosa, Peluzio, Telma Machado de Oliveira and Saito, Nathália Suemi.** *SPRING 5.1.2 passo a passo*. Alegre : INPE, 2010.
17. **Haykin, Simon.** *Neural Networks and Learning Machines*. 3ª. Upper Saddle River : Pearson Education, 2009. 8573077182.
18. **Dawson, C. W. and Wilby, R. L.** Hydrological modelling using artificial neural networks. *Progress in Physical Geography: Earth and Environment*. Março 1, 2001, Vol. 25, 1, pp. 80-108.
19. **Khan, Javed, et al.** Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*. Junho 2001, Vol. 7, pp. 673-679.

20. **Ludwig Jr., Oswaldo and Montgomery, Eduard.** *Redes Neurais*. Rio de Janeiro : Ciência Moderna, 2007. 9788573936193.
21. **Braga, Antônio de Pádua, Carvalho, André Ponce de Leon F. de and Ludermir, Teresa Bernarda.** *Redes Neurais Artificiais: Teoria e Aplicações*. Rio de Janeiro : LTC, 2000.
22. **Patterson, Josh and Gibson, Adam.** *Deep Learning: a Practitioners's Approach*. Sebastopol : O'Reilly, 2017.
23. **Eickenberg, Michael, et al.** Seeing it all: Convolutional network layers map the function of the human visual system. *Seeing it all: Convolutional network layers map the function of the human visual system*. Maio 15, 2017, Vol. 152, pp. 184-194.
24. **Zhang, Wei, et al.** Shift-Invariant Neural Network for Image Processing: Learning and Generalization. *Applications of Artificial Neural Networks III*. Setembro 16, 1992, Vol. 1709, pp. 257-268.
25. **Collobert, Ronan and Weston, Jason.** A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 160-167.
26. **Ferreira, Jonnison L., et al.** Segmentação Automática da Próstata em Imagens de Ressonância Magnética utilizando Redes Neurais Convolucionais e Mapa Probabilístico. *Simpósio Brasileiro de Computação Aplicada à Saúde*. Julho 26, 2018, Vol. 18, 1.
27. **Pereira, Sérgio, Pinto, Adriano, Alves, Victor and Silva, Carlos A.** Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Transactions on Medical Imaging*. Março 4, 2016, Vol. 35, 5, pp. 1240-1251.
28. **Ramachandran, Prajit, Zoph, Barret and Le, Quoc V.** Searching for Activation Functions. *CoRR*. 2017, Vol. abs/1710.05941.
29. **Szegedy, Christian, et al.** Going Deeper with Convolutions. *CoRR*. 2014.
30. **Simonyan, Karen and Zisserman, Andrew.** Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*. Maio 2014.
31. **Krizhevsky, Alex, Sutskever, Ilya and Hinton, Geoffrey E.** ImageNet classification with deep convolutional neural networks. *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*. NIPS'12, Dezembro 2012, Vol. 1, pp. 1097-1105.

32. **Ronneberger, Olaf, Fischer, Philipp and Brox, Thomas.** U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Maio 18, 2015, pp. 234-241.
33. **Badrinarayanan, Vijay, Kendall, Alex and Cipolla, Roberto.** SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Dezembro 2017, Vol. 39, 12, pp. 2481–2495.
34. **Yu, Fisher and Koltun, Vladlen.** Multi-Scale Context Aggregation by Dilated Convolutions. *International Conference on Learning Representation*. Abril 2016.
35. **Chen, Liang-Chieh, et al.** Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. Maio 2015, Vol. abs/1412.7062.
36. **Hasegawa, Akira, et al.** Convolution Neural Network based detection of lung structures. *Medical Imaging 1994: Image Processing*. Maio 11, 1994, Vol. 2167, pp. 654-662.
37. **Long, Jonathan, Shelhamer, Evan and Darrell, Trevor.** Fully Convolutional Networks for Semantic Segmentation. *CoRR*. Março 2015, 2015.
38. **Mathworks.** Semantic Segmentation Basics. *Matlab Documentation*. [Online] janeiro 7, 2019. [Cited: abril 2019, 23.] <https://www.mathworks.com/help/vision/ug/semantic-segmentation-basics.html>.
39. **LeCun, Y., et al.** Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*. Dezembro 1989, Vol. 1, 4, pp. 541-551.
40. **Gopalakrishnan, Kasthurirangan, et al.** Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection. *Construction and Building Materials*. Dezembro 30, 2017, Vol. 157, pp. 322-330.
41. **Liu, William Tse Horng.** *Aplicações de Sensoriamento Remoto*. 2ª. Campo Grande : UNIDERP, 2006.
42. **GeoHack.** GeoHack - São Luís (Maranhão). *GeoHack*. [Online] 2019. [Cited: março 5, 2019.] <https://bit.ly/2GmW0F5>.
43. **Russel, Stuart J. and Norvig, Peter.** *Artificial Intelligence: A Modern Approach*. 3rd. Upper Saddle River : Prentice Hall, 2009.

44. **Csurka, Gabriela, Larlus, Diane and Perronnin, Florent.** What is a good evaluation measure for semantic segmentation? *Proceedings of the British Machine Vision Conference*. 2013, pp. 32.1–32.11.
45. **Guérin, Joris, et al.** CNN features are also great at unsupervised classification. *4th International Conference on Artificial Intelligence and Applications*. 2017, Vol. abs/1707.01700.
46. **Kanezaki, Asako.** Unsupervised Image Segmentation by Backpropagation. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. Abril 2018, pp. 1543-1547.
47. **Plotze, Rodrigo de Oliveira and Bruno, Odemir Martinez.** Estudo e comparação de algoritmos de esqueletonização para imagens binárias. *IV Congresso Brasileiro de Computação*. 2004, pp. 59-64.
48. **IBGE.** Estado do Maranhão: Pedologia. *IBGE GeoFTP*. [Online] 2011. [Cited: dez 9, 2018.]
http://geoftp.ibge.gov.br/informacoes_ambientais/pedologia/mapas/unidades_da_federacao/ma_pedologia.pdf.
49. **Embrapa.** *Sistema brasileiro de classificação de solos*. 2ª. Brasília : Embrapa, 2006.
50. **Madeira Netto, José da Silva and Baptista, Gustavo Macedo de Mello.** *Reflectância espectral de solo*. Planaltina : Embrapa, 2000.
51. **Furht, Borko, Akar, Esad and Andrews, Whitney Angelica.** *Digital Image Processing: Practical Approach*. Cham : Springer, 2018.
52. **Skiena, Steven.** *Calculated Bets*. Cambridge : Cambridge University Press, 2004.
53. **Ferreira, Alessandro dos Santos.** *Redes Neurais Convolucionais Profundas na detecção de plantas daninhas em lavoura de soja*. Universidade Federal de Mato Grosso do Sul. Campo Grande : s.n., 2017. p. 80, Dissertação de Mestrado.
54. **Srinivas, Suraj, et al.** A Taxonomy of Deep Convolutional Neural Nets for Computer Vision. *Frontiers in Robotics and AI*. 2016, Vol. 2.