UNIVERSIDADE FEDERAL DE OURO PRETO
ESCOLA DE MINAS
COLEGIADO DO CURSO DE ENGENHARIA DE CONTROLE E AUTOMAÇÃO - CECAU

LEONARDO CARDOSO DA CUNHA

REDES NEURAIS CONVOLUCIONAIS E SEGMENTAÇÃO DE IMAGENS – UMA REVISÃO BIBLIOGRÁFICA

MONOGRAFIA DE GRADUAÇÃO EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Ouro Preto, 2020

LEONARDO CARDOSO DA CUNHA

REDES NEURAIS CONVOLUCIONAIS E SEGMENTAÇÃO DE IMAGENS – UMA REVISÃO BIBLIOGRÁFICA

Monografia apresentada ao Curso de Engenharia de Controle e Automação da Universidade Federal de Ouro Preto como parte dos requisitos para a obtenção do Grau de Engenheiro de Controle e Automação.

Orientadora: Profª. Luciana Gomes Castanheira.

Ouro Preto
Escola de Minas – UFOP
10/2020

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

C972r Cunha, Leonardo Cardoso da.
Redes Neurais Convolucionais e Segmentação de Imagens [manuscrito]: Uma revisão bibliográfica. / Leonardo Cardoso da Cunha. - 2020.
52 f.: il.: color..

Orientadora: Profa. Dra. Luciana Gomes Castanheira.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Escola de Minas. Graduação em Engenharia de Controle e Automação .

1. Redes Neurais (Computação) - Deep Learning. 2. Imagens - Segmentação de Imagens. 3. Imagens - Segmentação Semântica. 4. Redes Neurais (Computação) - Convolucionais. I. Castanheira, Luciana Gomes. II. Universidade Federal de Ouro Preto. III. Título.

CDU 681.5

Bibliotecário(a) Responsável: Maristela Sanches Lima Mesquita - CRB: 1716

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
DEPARTAMENTO DE ENGENHARIA CONTROLE E AUTOMACAO

FOLHA DE APROVAÇÃO

Leonardo Cardoso da Cunha

Redes Neurais Convolucionais e sua Aplicação em Segmentação de Imagens

Membros da banca

Agnaldo J Rocha Reis - Doutor - Universidade Federal de Ouro Preto
André Almeida Santos - Mestrando - Instituto Tecnológico Vale
Luciana Gomes Castanheira - Doutora - Universidade Federal de Ouro Preto

Versão final
Aprovado em 04 de dezembro de 2020.

De acordo

Professora Luciana Gomes Castanheira.

Documento assinado eletronicamente por **Luciana Gomes Castanheira**, **COORDENADOR(A) DO CURSO DE ENGENHARIA DE CONTROLE E AUTOMACAO**, em 18/01/2021, às 16:17, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0109112** e o código CRC **48A7B63D**.

Referência: Caso responda este documento, indicar expressamente o Processo nº 23109.009200/2020-77 SEI nº 0109112

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: 3135591533 - www.ufop.br

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

AGRADECIMENTOS

À Universidade Federal de Ouro Preto e seu corpo docente, pela oportunidade de realizar o curso em um ambiente produtivo e amigável.

Ao pesquisador André Almeida Santos pela elucidação de dúvidas técnicas do tema estudado e correções finais.

À minha orientadora Luciana, pelas suas correções, esclarecimentos e incentivos.

Aos meus pais e irmã, pelo apoio e por tornar possível a realização do curso.

RESUMO

No mundo atual, o uso de redes neurais para interpretar dados é cada vez mais comum. Redes Convolucionais são um tipo especializado de rede neural, capaz de interpretar imagens como dados de entrada. Essa poderosa ferramenta, é tradicionalmente aplicada em problemas de classificação, sendo utilizada por grandes empresas. Sua existência é devida entre outros à uma série de contribuições feitas pela comunidade científica ao longo das últimas décadas. Hoje, através de pesquisa bibliográfica, é possível conhecer a sua história e compreender a função das diferentes camadas que compõem essas redes: camadas de convolução, ReLU, *pooling* e as camadas totalmente conectadas. Essas últimas, num passado recente, ao serem substituídas por outra camada de convolução, deram origem às Redes Totalmente Convolucionais, capazes de segmentar imagens no estado da arte, possibilitando novas aplicações, como na visão de carros autônomos.

Palavras-chaves: Visão Computacional. *Deep-Learning*. Redes Neurais Convolucionais. Redes Totalmente Convolucionais. Segmentação semântica. Segmentação de imagens.

ABSTRACT

The usage of neural networks to interpret data has become increasingly common in the past years. Convolutional Neural Network is a particular class of neural network capable of interpreting visual imagery. This powerful tool, traditionally applied to classification events, has been adopted by large corporations as a result of contributions from the scientific community. Through bibliographic research, it is possible to comprehend its history and understand the function of the different layers that compose these networks: convolution, ReLU, pooling and fully connected layers. Recently, researchers have identified a new structure called Fully Convolutional Networks by replacing fully connected layers by another layer of convolution. This new architecture provides the means for state-of-the-art image segmentation, enabling new applications such as machine vision for driverless cars.

Keywords: Computer Vision. Deep Learning. Convolutional Neural Networks. Fully Convolutional Networks. Semantic segmentation. Image segmentation.

LISTAS DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

CNN – *Convolutional Neural Networks.*

FCN – *Fully Convolutional Networks.*

FC – *Fully Connected.*

CUDA - *Compute Unified Device Architecture.*

ILSVRC - *ImageNet Large Scale Visual Recognition Challenge*.

SUMÁRIO

1 INTRODUÇÃO

Redes Neurais Artificiais são sistemas de computação com nós interconectados inspirados nos neurônios do cérebro humano. Usando algoritmos, ou seja, linhas de código que processam e manipulam variáveis ordenadamente, essas redes podem reconhecer padrões escondidos e correlações em dados brutos. A partir de treinamento, essas redes podem aprender a modelar relações entre entradas e saídas (*input* e *output*) de dados complexos.

Uma rede neural pode ser projetada para classificar um objeto ao analisar suas características. Neste caso, deseja-se que a máquina seja capaz de categorizar o objeto como uma entre *k* classes (GOODFELLOW; BENGIO; COURVILLE, 2016). Por exemplo: uma rede pode ser treinada para prever se um cliente (*input*) deixará ou não o banco (*output*) no próximo mês, analisando seu saldo, emprego, idade, entre outras características desse cliente.

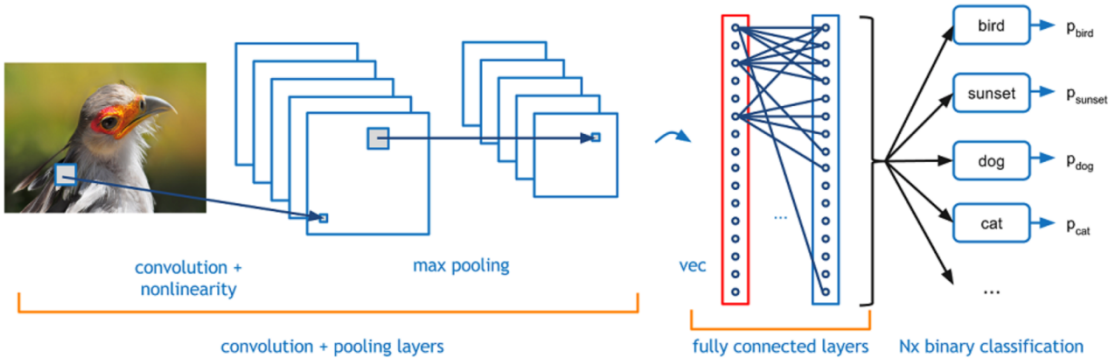
As redes aprendem essas relações através de treinamento: numa rede neural, para cada interação entre dois neurônios é atribuído um peso, um fator multiplicativo usado em cálculos internos. O treinamento consiste em ajustar esses pesos, a fim de reduzir o erro observado ao comparar a saída da rede com o resultado esperado. Para isso, um banco de dados de entrada e gabarito são utilizados (treinamento supervisionado). A qualidade e quantidade dos dados utilizados nessa fase irá impactar diretamente no treinamento e consequente precisão da rede (GOODFELLOW; BENGIO; COURVILLE, 2016).

O que acontece, entretanto, quando se deseja abordar imagens como objetos de entrada? A imagem é um tipo especial de dado, que pode ser pensado como uma grade em 2D de *pixels*. Uma classe especializada de rede neural tem experimentado um tremendo sucesso em aplicações práticas envolvendo imagens, são as chamadas Redes Neurais Convolucionais (CNN do inglês) (GOODFELLOW; BENGIO; COURVILLE, 2016).

Essas redes possuem uma estrutura comumente segmentada em camadas, entre elas estão as chamadas camadas de convolução. Essas são responsáveis pela identificação de características da imagem. Elas podem ser pensadas como filtros que irão percorrer a imagem, cada uma em busca de uma característica específica. Entende-se por características: bordas retas, curvaturas específicas, cores simples etc. Ou seja, atributos abstratos que toda imagem tem em comum (DESHPANDE, 2016).

A ideia é que uma camada de rede neural totalmente conectada classifique a imagem a partir das características identificadas e mapeadas nas camadas de convolução conforme a figura 1 (DESHPANDE, 2016).

Figura 1 - Camadas de uma Rede Neural Convolucional para classificação.



Extraído de: DESHPANDE, 2016

As CNN's foram inspiradas no córtex visual, e são, portanto, uma combinação de biologia, matemática e computação - o que pode soar estranho à primeira vista. Todavia, as CNN's demonstraram ser uma ferramenta poderosa e vêm sendo usadas por grandes empresas em seus serviços: *Google* na pesquisa de fotos, *Amazon* na recomendação de produtos, *Instagram* na estrutura de pesquisa, *Facebook* no sistema de marcação automática, entre outros. (DESHPANDE, 2016).

Não há dúvidas que as CNN's são poderosos instrumentos. Entretanto, em determinadas aplicações, é possível apontar limitações nessas redes. Elas não são capazes, por exemplo, de classificar mais de um tipo de objeto em uma mesma figura, nem de localizar objetos dentro da imagem. Essas propriedades podem ser essenciais numa aplicação que envolve maior grau de autonomia, como na visão de carros autônomos.

Neste contexto, as redes totalmente convolucionais (FCN do inglês) se mostraram uma alternativa inovadora. Nesta abordagem, que é uma variante das CNN's clássicas, ao invés de uma única distribuição de probabilidades para a imagem como um todo, cada *pixel* tem a sua própria distribuição de probabilidades. A rede é treinada para prever a qual classe o *pixel* pertence, e retorna uma imagem segmentada por classes (figura 2), permitindo-a classificar vários tipos de objetos, e localizá-los (HEINRICH, 2016; LONG; SHELHAMER; DARRELL, 2017).

Figura 2 - Imagem segmentada por classes retirada do banco de imagens PASCAL VOC.

Extraído de: HEINRICH, 2016

O presente trabalho abrange a histórica da evolução dos conhecimentos científicos acerca das redes neurais de convolução ao longo dos anos, e descreve os mecanismos e funções por trás de cada camada que compõe sua estrutura básica. Por fim, explica como são formadas as redes para segmentação semântica de imagens.

1.1 Justificativa

As redes neurais de convolução são uma tecnologia recente e promissora. Sua flexibilidade permite diversos tipos de aplicação em diferentes áreas. Entender seu funcionamento através de estudo e publicações acadêmicas torna-se importante, conhecendo assim sua fundamentação teórica respaldada na ciência.

1.2 Objetivo geral

Descrever a evolução e os componentes básicos das Redes Neurais Convolucionais (CNN) até a inserção das Redes Totalmente Convolucionais (FCN), possibilitando a segmentação semântica de imagens.

1.3 Objetivos específicos

Como objetivos específicos do trabalho, tem-se a descrição:

• Dos componentes básicos de uma rede neural de convolução e o processo de segmentação semântica de imagens a partir dela.

• Da evolução histórica do conhecimento científico sobre as redes neurais de convolução até o desenvolvimento das redes totalmente convolucionais.

2 METODOLOGIA

O trabalho desenvolvido seguiu os preceitos do estudo exploratório, por meio de uma pesquisa bibliográfica, que, segundo Gil (2008, p. 50), “é desenvolvida a partir de material já elaborado, constituído de livros e artigos científicos”.

Foram utilizados os livros *Deep Learning* de Goodfellow, Bengio e Courville, e *The Organization Of Behavior,* escrito por Donald Hebb em 1949.

Artigos científicos sobre a temática foram pesquisados nas bases de dados: Semantic Scholar e IEEE Xplore. Sendo em alguns casos acessados através do arXiv. Publicados nos últimos 31 anos (1989 – 2020). Foram utilizados os seguintes descritores: *convolutional networks, cnn survey, cnn architectures, semantic segmentation, understanding cnn.* O levantamento realizado nas bases selecionadas teve por critério de busca os “mais citados”. Em alguns casos, foi utilizado “mais relevantes” ou “com maior influência”.

O blog de Adit Deshpande, estudante de ciência da computação da Universidade da Califórnia em Los Angeles, forneceu muitas informações relevantes para o desenvolvimento do trabalho. O autor teve 8 postagens republicadas, sendo 4 delas premiadas pelo *KDnuggets*: site várias vezes premiado, com numerosas menções como líder em publicações em Inteligência Artificial.

Outros blogs, devidamente referenciados, foram usados quase exclusivamente para a extração de imagens, e foram encontrados através da ferramenta de pesquisa de imagens do *Google*.

2.1 Coleta de dados

A coleta de dados seguiu as seguintes etapas:

a) Leitura exploratória de todo o material selecionado (leitura rápida que visa verificar se a obra é de interesse para o trabalho).
b) Leitura seletiva (leitura mais aprofundada das partes que interessam o estudo).
c) Registro das informações extraídas (autores, ano e resultados).

2.2 Análise dos resultados

Foi realizada uma leitura analítica que possibilitasse o entendimento do avanço do conhecimento científico do tema pesquisado em ordem cronológica, a fim de ordenar e sumariar as informações contidas nas fontes.

3 REVISÃO BIBLIOGRÁFICA

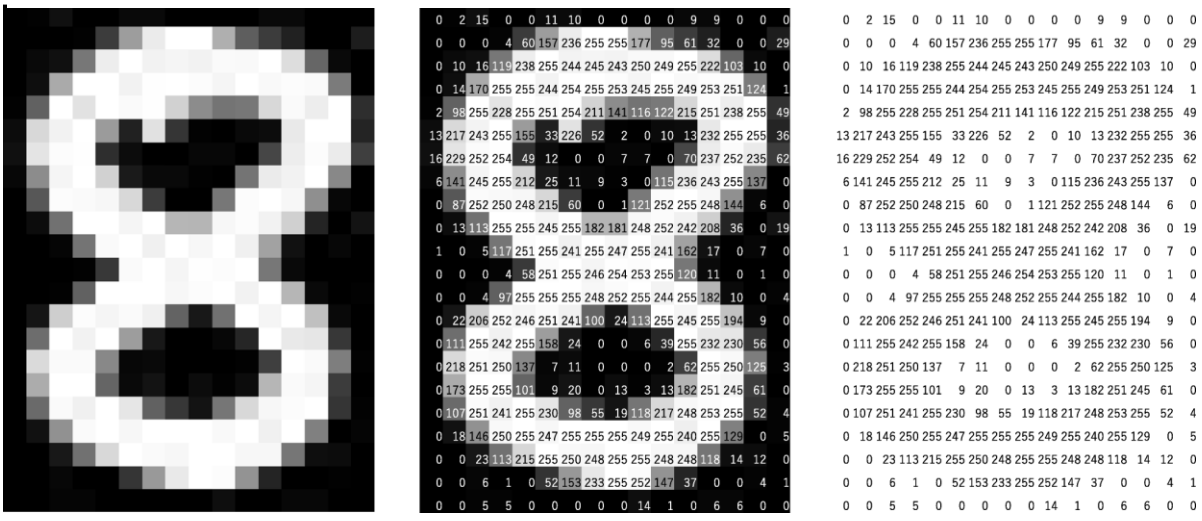
3.1 O propósito das redes de convolução

3.1.1 Como as máquinas “enxergam”?

Quando os seres humanos olham para uma imagem, ou para o mundo a sua volta, são capazes de identificar cenários e nomear objetos sem muito esforço. Isso é feito inconscientemente. Essa habilidade de reconhecer padrões rapidamente, generalizar e adaptar ambientes em diferentes situações é uma habilidade que, diferente dos humanos, as máquinas não têm (DESHPANDE, 2016).

Quando um computador recebe uma imagem como entrada, ele se depara com uma série de números, que representam os *pixels*. A máquina irá organizar esses *pixels* em forma de matriz. Cada número da matriz tem um valor de 0 a 255 que representa a intensidade de cor do *pixel*, conforme apresentado na figura 3. Esses números, apesar de sem muito significado para nós humanos, são o único material disponível para as máquinas. É assim que elas “enxergam” as imagens (DESHPANDE, 2016).

Figura 3 - A imagem como matriz de *pixels*.



Extraído de: ÜNAL, 2019

3.1.2 O que é a tarefa de classificação de imagens?

O que se deseja é que essas máquinas sejam capazes de diferenciar dentre as imagens dadas, as características únicas que faz de um cachorro um cachorro, ou que faz de um gato um gato. Numa abordagem geral, pode-se dizer que as redes neurais de convolução irão procurar por formatos característicos em baixo nível da imagem, como curvaturas, bordas e pontas. E ao conduzir essas características por uma série de camadas com diferentes funções, a rede irá retornar qual é a classe que melhor descreve a imagem, gerando uma divisão de probabilidades (figura 4) (DESHPANDE, 2016).

Figura 4 - Classificação da rede Alexnet da imagem de um gato do banco de imagens PASCAL VOC.

Extraído de: HEINRICH, 2016.

3.1.3 O que é a tarefa de segmentação semântica de imagens?

Além da tarefa de classificação de imagem, onde uma única descrição em alto nível é atribuída à imagem, recentemente pesquisadores descobriram que ao modificar sua estrutura, as redes de convolução são capazes de realizar outras tarefas no estado da arte, como detecção de objetos, segmentação semântica de imagens e segmentação de instâncias (figura 5), fornecendo informações mais detalhadas e localizadas no ambiente (ARNAB *et al*, 2018).

Figura 5 - Tarefas de detecção de objetos (esquerda), segmentação semântica (centro), segmentação semântica com diferenciação de instâncias (direita).

Extraído de: ARNAB *et al*, 2018

No caso da segmentação semântica, o objetivo é compreender melhor a cena ao atribuir a cada *pixel* no interior da imagem, uma categoria de objeto. Essa compreensão permite que as máquinas extraiam informações de diferentes cenários do mundo real, possibilitando a realização de diferentes tarefas (ARNAB *et al.*, 2018).

Destacam-se entre as aplicações de segmentação semântica de imagens: nos veículos autônomos, onde é necessário um entendimento preciso do ambiente; no diagnóstico médico, onde células, tecidos e órgão de interesse são segmentados; no desenvolvimento de robôs que navegam e manipulam objetos no ambiente; na edição de imagens e vídeos; e no desenvolvimento de “óculos-inteligentes” que descrevem a cena para deficientes visuais (ARNAB *et al.*, 2018).

3.2 Inspiração neurobiológica: o córtex visual

Atualmente, as CNN’s são consideradas os algoritmos mais amplamente utilizados entre as técnicas de inteligência artificial inspiradas na biologia (GOODFELLOW; BENGIO; COURVILLE, 2016). Sua história começou com experimentos neurobiológicos conduzidos por David Hubel e Torsten Wiesel.

A arquitetura neuronal como objeto de estudo começou a ser estimulada em 1949, ano em que Donald Hebb (1949) publicou o livro *The Organization of Behavior*, onde apresentou suas ideias sobre o comportamento neuronal baseado na organização de circuitos neuronais. Suas ideias foram amplamente lidas, e apesar de

seu estudo ter sido baseado em neurônios hipotéticos, Hebb incentivou o pensamento sobre como os neurônios poderiam ser organizados para produzir comportamentos complexos.

Anos mais tarde, Hubel & Wiesel (1959, 1962), fundamentaram o entendimento do sistema visual com experimentos que se estenderam por mais de 25 anos. Sua primeira publicação em 1959 e subsequente extensão em 1962 fizeram um marco na exploração de como os neurônios do cérebro podem ser organizados para produzir percepção visual.

Em seus experimentos, Hubel e Wiesel basearam-se em observações anteriores sobre o mecanismo de entrada no córtex visual, a retina. Um preciso trabalho escrito por Stephen Kuffler (1953) identifica células ganglionares na retina de gatos, e a organização entre elas. Eles construíram o procedimento descrito por Kuffler para encontrar os estímulos necessários para ativação de cada célula. Com os experimentos, mostraram que apesar de pontos de luz serem os estímulos mais eficientes para ativar as células ganglionares da retina, as fendas de luz orientadas estimulavam mais os neurônios do córtex. Assim, eles demonstraram que a diferença do estímulo preferido pelo córtex em comparação ao da retina é devido à resposta cortical sendo derivada de um nível baixo, e gradativamente aumentando de complexidade ao longo de camadas dentro do córtex. Além da possível contribuição de regiões excitatórias e inibidoras do campo de visão circular dos neurônios de entrada.

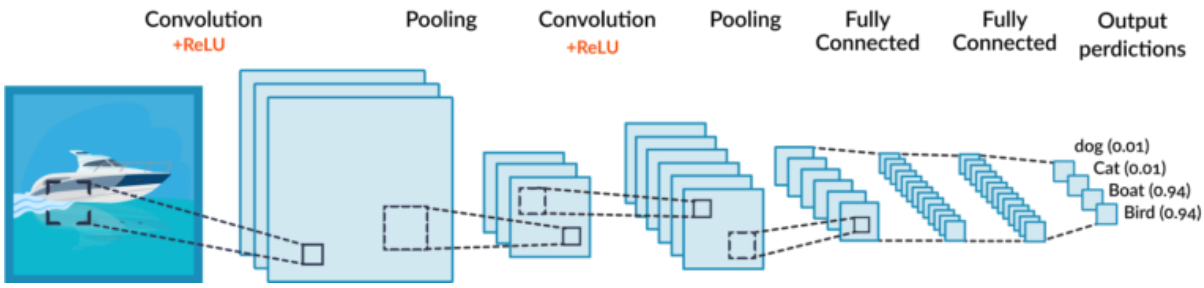
A compreensão de que as sucessivas transformações no córtex visual são resultado de uma organização celular arquitetada, serviu de inspiração na criação das redes neurais convolucionais. E levou ao questionamento do que acontece nos níveis seguintes e a natureza das transformações nas regiões mais altas do córtex (GOODFELLOW; BENGIO; COURVILLE, 2016).

3.3 Componentes básicos das redes de convolução

Conforme dito anteriormente, uma imagem passa por uma série de camadas dentro da rede e retorna uma saída, que pode ser uma divisão de probabilidades. No fim das contas, o arranjo dessas camadas e componentes têm um papel fundamental na criação de diferentes arquiteturas de rede, cada uma com suas vantagens e desvantagens (KHAN *et al*., 2020).

Tipicamente, a arquitetura básica de uma CNN é composta de camadas alternadas de convolução e *pooling*, seguidas por uma ou mais camadas totalmente conectadas (figura 6). Já adiantando o assunto do tópico 3.5: no caso das redes de convolução para segmentação de imagem, as camadas totalmente conectadas (FC) serão substituídas por uma camada final de convolução, sendo, portanto, chamadas de Redes Totalmente Convolucionais (FCN) (LONG; SHELHAMER; DARRELL, 2015).

Figura 6 - As camadas de uma CNN de classificação.



Extraído de: Página do Blog MissingLink[1].

3.3.1 Camada de convolução

Convolução é uma operação matemática entre duas funções para produzir uma terceira, que expressa o quanto o formato da primeira é modificado pela segunda (RIZWAN, 2018). Quando se trata de redes convolucionais, o primeiro argumento é a imagem de entrada (*input*), o segundo é chamado de filtro (em inglês o termo *kernel*

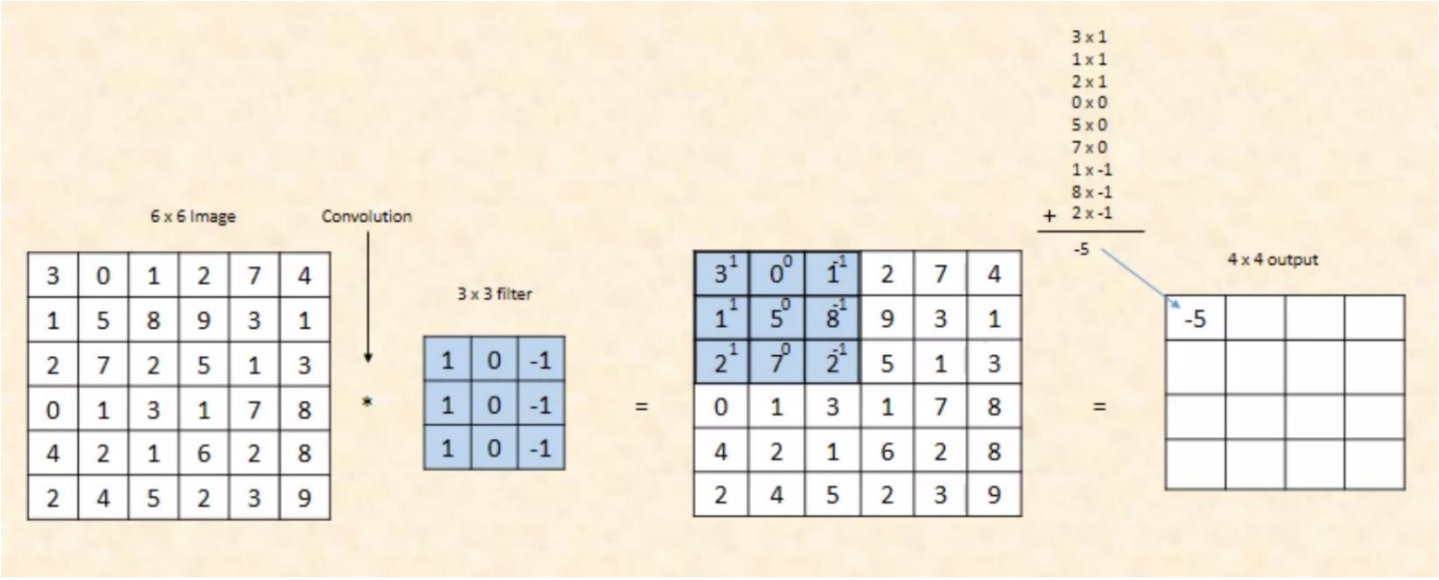
[1] Disponível em: https://missinglink.ai/guides/convolutional-neural-networks/convolutional-neural-network-tutorial-basic-advanced/ Acesso em: 10 jul. 2020.

é mais comum). Já a saída é referida como mapa de ativação (*feature map*) (GOODFELLOW; BENGIO; COURVILLE, 2016). Entender o conceito e a matemática por trás do processo de convolução é essencial, pois esse processo se repetirá múltiplas vezes em diferentes camadas da rede.

Imaginemos um feixe de luz que cobre um pedaço da imagem, agora, partir do topo esquerdo da imagem, esse feixe desliza percorrendo toda a imagem. Suponha que esse feixe de luz carregue consigo uma série de pesos, e ao passar por cada região da imagem, é gerado o produto entre os pesos do feixe e os valores dos *pixels* da região da imagem. Nessa analogia, o feixe de luz é o filtro, e a região que o filtro cobre é chamado de campo receptivo. Já os produtos gerados, são armazenados nos chamados mapas de ativação (DESHPANDE, 2016).

Um exemplo pode ser observado na figura 7. Para cada campo receptivo, é realizada a soma dos produtos obtidos, restando um único número. Esse número representa o quanto aquela região foi ativada pelo filtro.

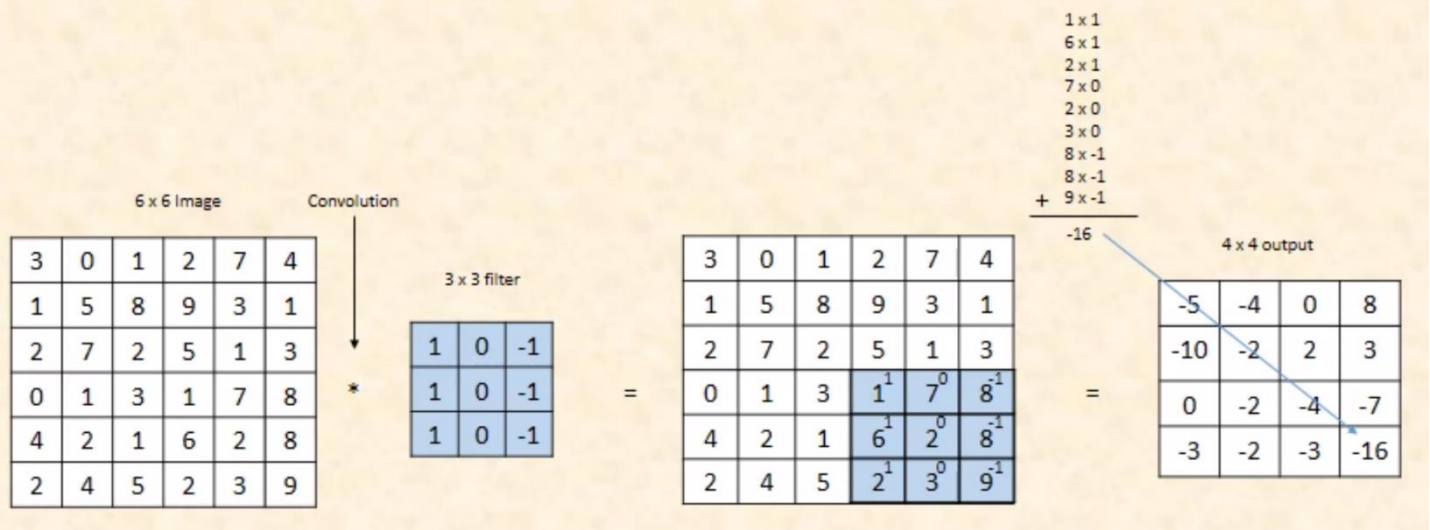
Figura 7 – Convolução entre imagem de entrada 6x6 e filtro 3x3, *stride* 1.



Extraído de: RIZWAN, 2018.

Esse processo se repete até que toda a imagem tenha sido percorrida, dando origem a um mapa de ativação (figura 8).

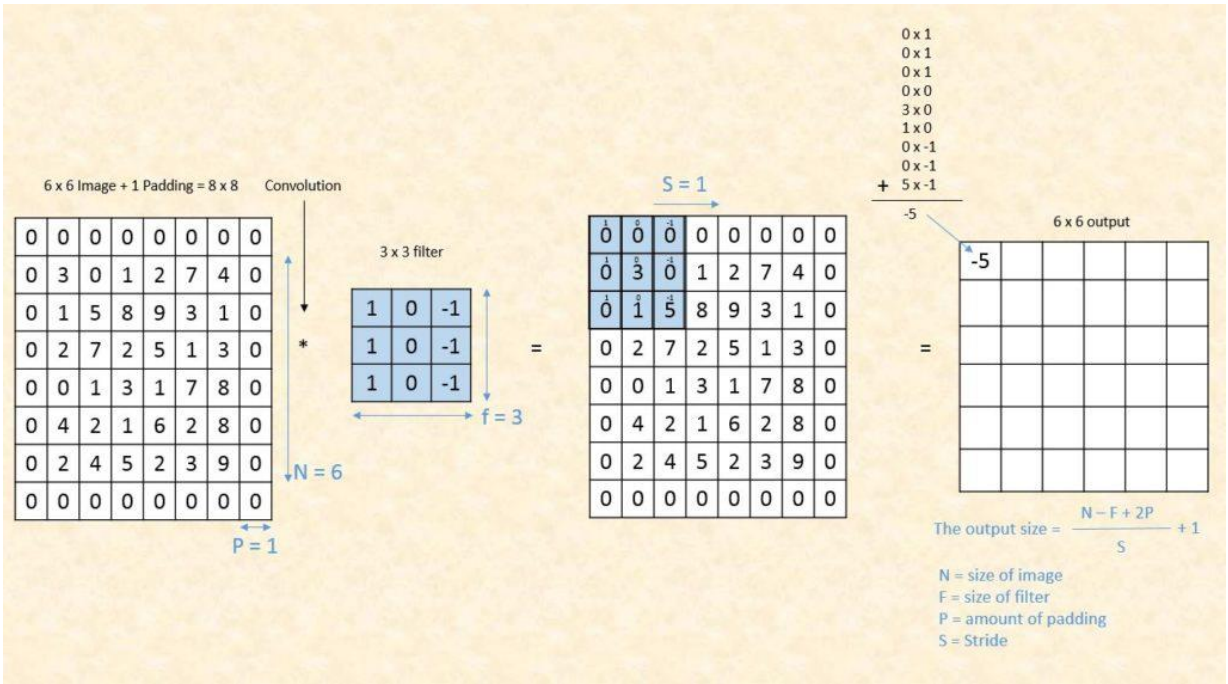
Figura 8 - Um mapa de ativação com dimensões 4x4 é gerado.



Extraído de: RIZWAN, 2018.

No exemplo, uma imagem de entrada com dimensões 6 x 6 e um filtro de tamanho 3 x 3 gerou um mapa de ativação de tamanho 4 x 4. Em alguns casos, esse encolhimento é indesejável. Nesse caso, adiciona-se à imagem original uma borda de zeros, chamada de *zero padding* (figura 9) (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 9 - Uma borda de zeros chamada *zero padding* impede que os dados encolham na convolução.



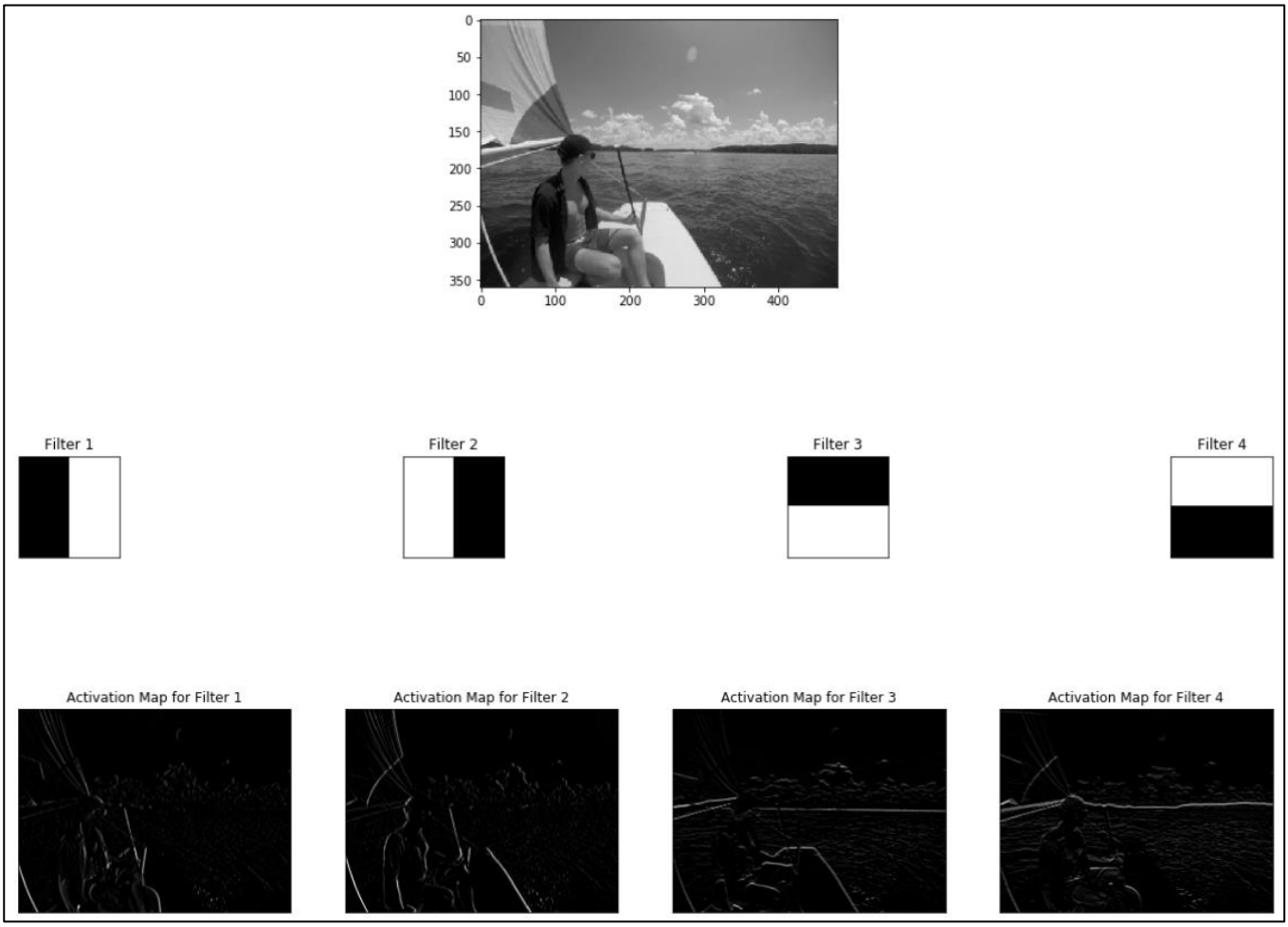
Extraído de: RIZWAN, 2018.

Já a distância que o filtro percorre entre um campo receptivo e outro é chamada de *stride*. No exemplo, o *stride* é igual a 1 pois o filtro percorre 1 *pixel* a cada interação. A combinação dessas variáveis permite controlar o tamanho do mapa de ativação gerado (RIZWAN, 2018).

Um mapa de ativação é gerado para cada filtro usado na camada de convolução, e representa as regiões onde as características do filtro foram encontradas (GOODFELLOW; BENGIO; COURVILLE, 2016).

Na figura 10, quatro diferentes filtros que detectam características de bordas foram aplicados à uma imagem, gerando quatro mapas de ativação.

Figura 10 – Mapas de ativação gerados usando quatro filtros.

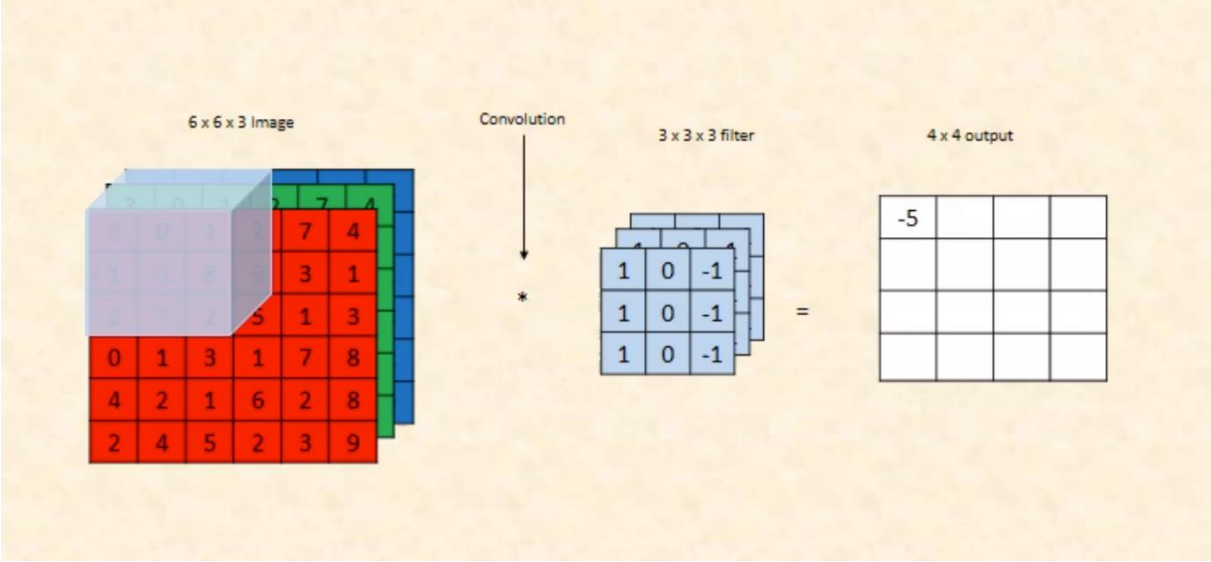


Extraído de: JORDAN, 2017

A diferença, no caso de imagens coloridas, é que a entrada possui três dimensões. Como pode-se observar no exemplo da figura 11, a entrada tem dimensões 6 x 6 x 3 onde o 3 denota o número de canais, que representam as intensidades de vermelho, verde e azul de cada *pixel* (RGB). É assim que as imagens coloridas são representadas.

Neste caso, a operação matemática é similar. A diferença está no filtro, que também deve conter 3 canais. Após a soma dos produtos entre campo receptivo e respectivo canal de filtro, soma-se os 3 valores encontrados, restando novamente um único número.

Figura 11 - Os canais RGB de uma imagem colorida.

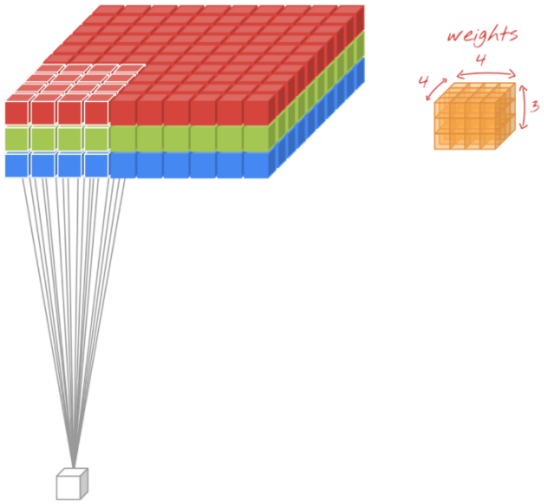


Extraído de: RIZWAN, 2018.

É importante notar que a largura e comprimento do filtro pode variar da maneira que se desejar, entretanto, para que a matemática funcione, a profundidade do filtro deve acompanhar a profundidade da entrada (mesmo número de canais) (DESHPANDE, 2016).

Na figura 12, tem-se uma representação em 3D de uma imagem colorida e um filtro 4 x 4.

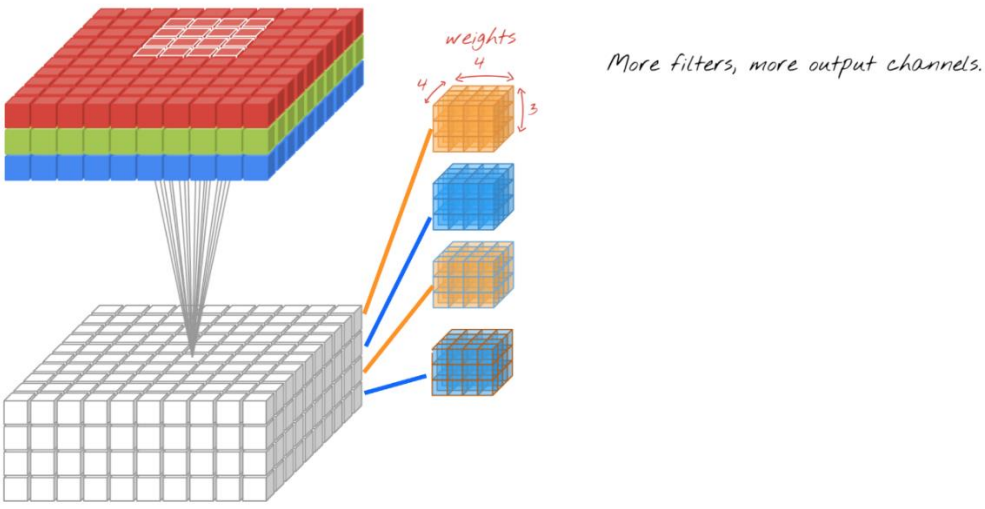
Figura 12 - Representação tridimesional de imagem e filtro.



Extraído de: GÖRNER, 2018

Uma pilha de mapas de ativação também pode ser referido como bloco de dados, uma vez que tem altura, largura e profundidade (figura 13).

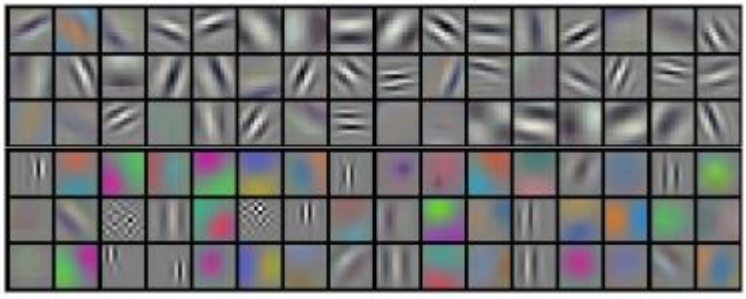
Figura 13 - Uma pilha de mapas de ativação sendo gerado por uma série de filtros.



Extraído de: GÖRNER, 2018

Vale destacar que não é papel do projetista da rede definir explicitamente os filtros que serão utilizados. Ao invés disso, esses filtros são apenas hiperparametrizados (isso significa que serão especificados ao algoritmo algumas medidas e atributos necessários para realizar o treinamento da rede, como tamanho e quantidade de filtros por camada, *stride, padding*, etc), e a rede irá aprender por conta própria os filtros que extraem características da imagem (figura 14) (JORDAN, 2017).

Figura 14 – 96 filtros tamanho 11x11x3 da primeira camada de convolução de uma rede de convolução.



Extraído de: KRIZHEVSKY; SUTSKEVER; HINTON, 2017

Numa segunda camada de convolução da rede, as entradas passam a ser os mapas de ativação gerados na primeira camada. Agora o filtro terá profundidade igual ao número de mapas de ativação gerado na camada anterior, garantindo que a matemática funcione.

Visualizar a função da segunda camada pode ser menos intuitivo quando comparado a primeira, afinal estamos convolvendo não mais uma imagem original que nós humanos vemos e compreendemos com facilidade, estamos agora aplicando convolução numa pilha de mapas de ativação gerados pela camada anterior.

O que acontece, é que os mapas de ativação gerados na primeira camada estão descrevendo as regiões onde determinadas características de baixo nível foram encontradas (como retas, pontas, bordas, e a presença de cores). E ao aplicar uma camada de convolução a esse bloco de dados, as saídas serão ativações que representam características de nível mais alto como quadrados (combinação de várias linhas retas), semicírculos (combinação de uma curva e uma reta), etc. De maneira

geral, à medida que se aprofunda na rede, passando por cada vez mais camadas de convolução, os mapas de ativação passam a representar características cada vez mais abstratas da imagem (ZEILER; FERGUS, 2014, DESHPANDE, 2016).

Pode-se, inclusive, constatar que conforme se aprofunda na rede, é muito comum ter um aumento gradativo no número de mapas de ativação. Isso é propositalmente projetado, pois uma vez que os mapas de ativação ficam cada vez mais especializados, representando conceitos mais abstratos, simplesmente são necessários mais mapas de ativação para representar a imagem original (JORDAN, 2017).

Agora uma observação: analisando o processo de convolução, é possível perceber que os pesos de um filtro são reutilizados, afinal um mesmo filtro percorre toda a imagem. Essa é uma importante propriedade das CNNs: é o chamado compartilhamento de parâmetros. Isso garante um número drasticamente menor de parâmetros únicos, e aumenta o tamanho da rede sem que seja necessário aumentar a quantidade de dados de treinamento (GOODFELLOW; BENGIO; COURVILLE, 2016).

3.3.2 Camada não-linear

Após as camadas de convolução, é convenção aplicar uma camada não-linear (ou camada de ativação) logo em seguida. O propósito desta camada é introduzir não-linearidade a um sistema que basicamente só tem computado operações lineares na camada de convolução (multiplicações e somatórios) (GOODFELLOW; BENGIO; COURVILLE, 2016).

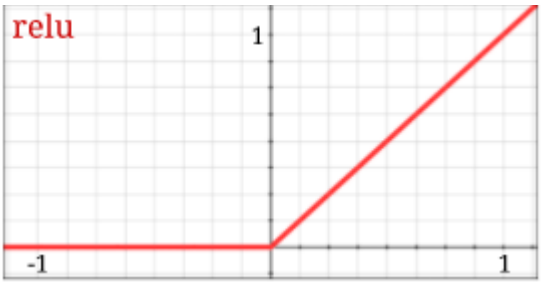
Introduzir não-linearidade a um modelo, torna-o capaz de criar associações mais complexas entre entradas e saídas da rede, o que é essencial para aprendizado de dados complexos, como imagens (GOODFELLOW; BENGIO; COURVILLE, 2016).

No passado, funções não-lineares como tangente hiperbólicas e sigmoides eram usadas como funções não-lineares, mas pesquisas recentes mostraram que a

função ReLU (do inglês *Rectified Linear Units*) traz melhores resultados pois permite que a rede seja treinada muito mais rapidamente (devido à eficiência computacional) sem diferenças significativas na precisão (KRIZHEVSKY; SUTSKEVER; HINTON, 2017; DESHPANDE, 2016).

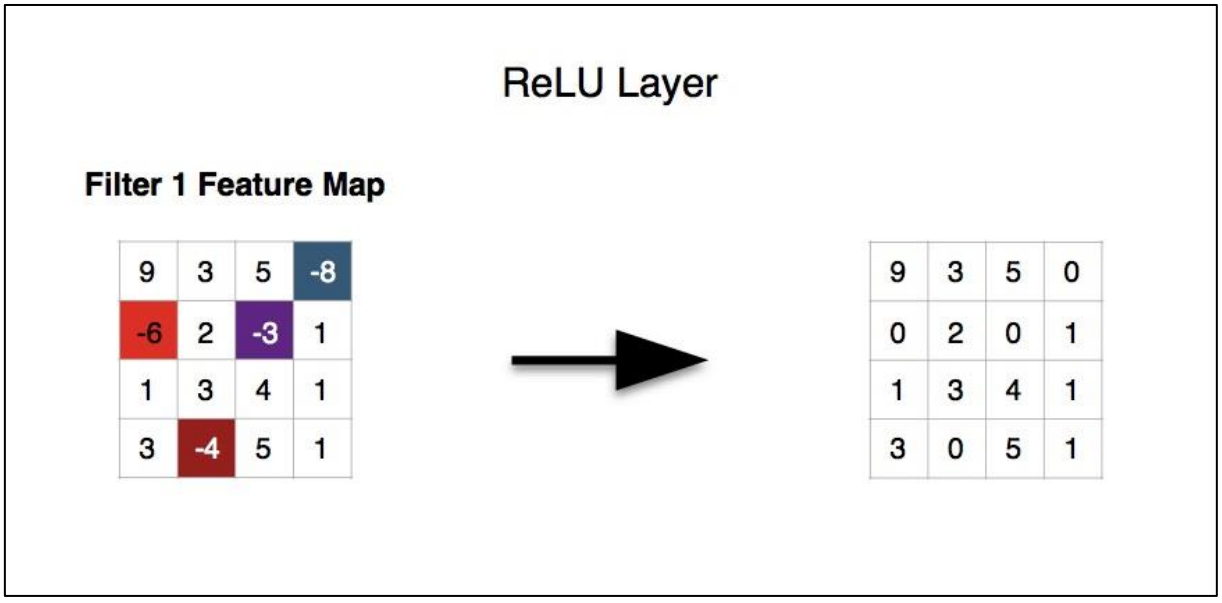
A camada ReLU aplica a função f(x) = max(0, x) a todos os valores do volume de entrada (figura 15). Em outras palavras, essa camada basicamente troca todos os valores negativos de ativação na convolução, em 0 (figura 16). Essa camada aumenta as propriedades não-lineares do modelo sem afetar as ativações da camada de convolução.

Figura 15 - Função ReLU.



Extraído de: GÖRNER, 2018.

Figura 16 - Camada ReLU em um mapa de ativação.



Extraído de: ANIEMEKA, 2017.

3.3.3 Camada de *pooling*

A fim de aumentar a quantidade de camadas da rede, é necessário condensar o tamanho espacial da representação da imagem. A ideia da camada de *pooling* é reduzir drasticamente o número de parâmetros. O processo, que pode ser chamado de *downsampling,* permite acelerar os cálculos computacionais da rede (RANZATO *et al*., 2007).

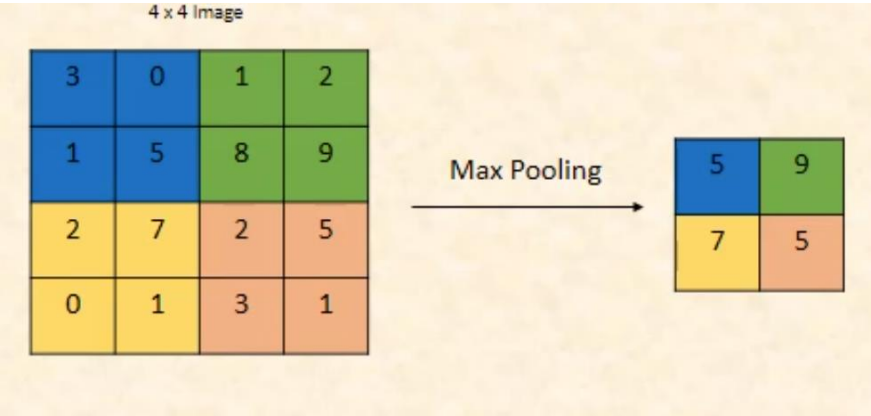
Outro importante benefício dessa camada é o aumento da invariância espacial da rede. Isso significa que a rede ficará mais robusta à possíveis torções, achatamentos ou qualquer outra variação na posição das características mapeadas (RANZATO *et al*., 2007).

É fácil entender seu mecanismo: de maneira similar à convolução, define-se um tamanho retangular de área e um *stride*, mas desta vez, ao invés de extrair informações, o objetivo é comprimir as informações já extraídas. A seguir estão os dois tipos mais utilizados de *pooling*.

3.3.3.1 *Max pooling*

Nessa abordagem, o valor máximo no interior do retângulo é retornado para cada região percorrida (figura 17). É o mais comum entre os tipos de *pooling* (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 17 - *Max pooling* retângulo 2x2 e *stride* 2.



Extraído de: RIZWAN, 2018

3.3.3.2 *Average pooling*

Nessa outra abordagem, a média dos valores no interior do retângulo é retornado para cada região percorrida (figura 18) (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 18 - *Average pooling* retângulo 2x2 e *stride* 2.

Extraído de: RIZWAN, 2018.

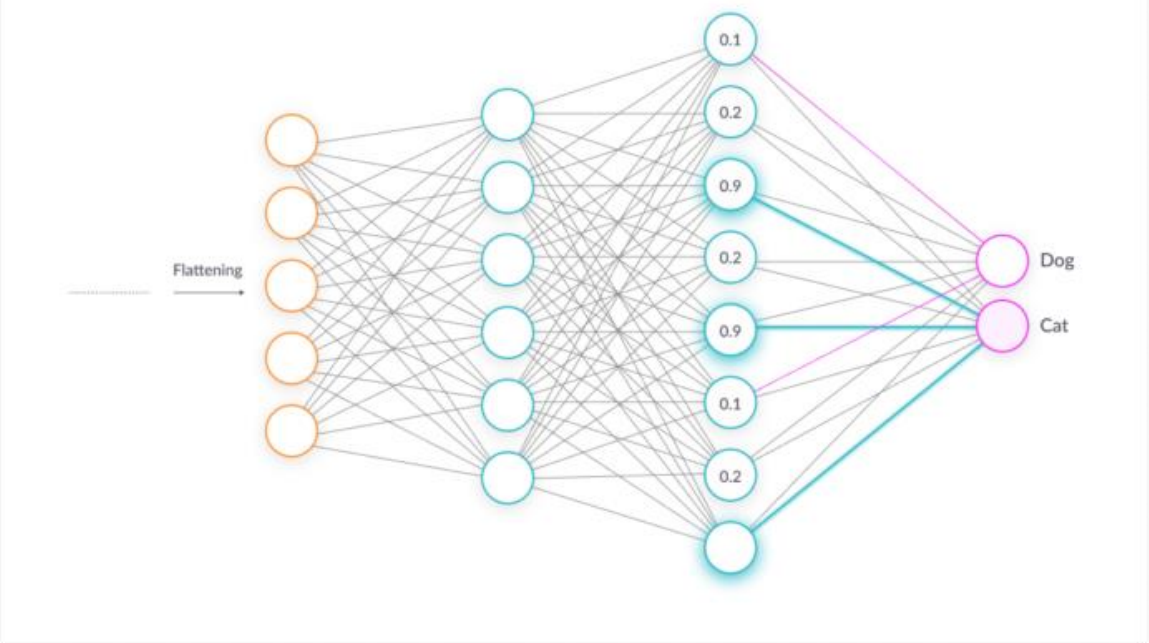
3.3.4 Camada totalmente conectada

A camada totalmente conectada de uma CNN cumpre o papel de retornar uma decisão quanto a classificação da imagem como um todo. E é, portanto, parte da estrutura das redes de classificação. Na realidade, o que torna possível essa classificação é o mecanismo de análise e extração de características isoladas da imagem, provenientes das camadas anteriores (convolução e *pooling*) (DESHPANDE, 2016).

A camada totalmente conectada irá interligar todas as características e atributos extraídos da imagem através de uma estrutura neuronal tradicional de

camadas ocultas, entretanto, onde todos os neurônios da camada estão interconectados (figura 19).

Figura 19 – Representação das camadas totalmente conectadas de uma CNN.



Extraído de: Página do Blog MissingLink[2].

A saída da camada anterior deve passar por um processo de achatamento (*flattening*) antes de entrar na camada totalmente conectada: os mapas de ativação que até então eram representados como matrizes, são transformados em um único vetor que irá alimentar a camada totalmente conectada (JORDAN, 2017).

Ao fazer uma análise das saídas da camada anterior (que devem representar os mapas de ativações de características de alto nível de abstração) a camada totalmente conectada determina quais classes mais se correlacionam com determinadas características. Por exemplo, se a rede está classificando uma imagem como cão, ela certamente obteve altos valores nos mapas de ativação que representam características de pata, ou de 4 pernas. Similarmente, se a rede está prevendo a classe pássaro para uma determinada imagem, isso ocorre pois os mapas de ativação que representam bicos ou asas obtiveram valor alto (DESHPANDE, 2016). A rede aprende a relacionar essas características a partir dos treinamentos a

[2] Disponível em: https://missinglink.ai/guides/convolutional-neural-networks/fully-connected-layers-convolutional-neural-networks-complete-guide/ Acesso em: 1 set. 2020

que são submetidos. Uma função softmax pode ser usada para obter os valores das previsões em porcentagem.

3.4 Evolução histórica das redes de convolução

No decorrer das últimas décadas, uma série de esforços levaram ao avanço no entendimento e evolução das redes de convolução. A evolução histórica das CNN's até a descoberta das FCN's pode ser categorizada em 5 grandes eras, e são discutidas abaixo.

3.4.1 Origem das CNN's: 1989 – 1999

LeCun *et al*. (1989) apresentaram pela primeira vez uma CNN multicamadas. Chamada de ConvNet, essa rede fazia uso de um algoritmo de *backpropagation*, possibilitando a rede a aprender padrões de filtros automaticamente, diferenciando-se de redes anteriores (KHAN *et al*., 2020).

A ConvNet obteve resultados bem-sucedidos para problemas de reconhecimento de dígitos e CEP escritos a mão. Anos mais tarde, essa arquitetura foi melhorada: chamada de LeNet-5, essa rede já era capaz de reconhecer variantes rotacionais das imagens. Sua boa performance permitiu que fossem usadas comercialmente em caixas eletrônicos e em bancos, em 1993 e 1996 respectivamente (LECUN *et al*., 1998; KHAN *et al*., 2020).

3.4.2 Estagnação das CNN's: Início dos anos 2000

O período que se compreende entre o final da década de 90 e 2006, trouxe poucos avanços na área. O interesse em redes neurais foi reduzido e, portanto, pouca atenção foi dada às CNN's. Uma das razões disso, é que o custo computacional exigido para que essas redes fossem treinadas não eram compatíveis com o que se tinha na época, e acabaram perdendo espaço para outros métodos estatísticos como

as SVM's (do inglês*: support vector machine*). No início dos anos 2000, acreditava-se amplamente que algoritmos de *backpropagation* usados para treinar CNN's não eram eficientes para convergir a um ponto ótimo, e que, portanto, não eram capazes de aprender recursos tão bem quanto outros métodos de aprendizado supervisionado, como os manuais (KHAN *et al*., 2020).

No entanto, outros pesquisadores continuaram a estudar as CNN's na tentativa de otimizar sua arquitetura. Simard *et al.* (2003) aperfeçoaram a arquitetura das CNN's, mostrando bons resultados em comparação às SVM's numa base de dados de benchmark para classificar dígitos escritos a mão. O aprimoramento da rede, estendeu a até então restrita aplicabilidade das CNN's. Entre os campos explorados está o sistema de detecção de rosto, que mais tarde foi usado comercialmente em mercados para rastreamento de clientes (KHAN *et al*., 2020).

3.4.3 O renascimento das CNN's: 2006 - 2011

A fim de solucionar os problemas que freavam o aprendizado das redes profundas, dois trabalhos publicados em 2006 e 2007, tiveram imensa contribuição principalmente na reintegração da importância desse tipo de rede: Hinton *et al*. (2006) propôs utilizar o método *greedy layer-wise unsipervised training*. Ranzato *et al*. (2007) usaram *max pooling*, mostrando bons resultados no aprendizado e introduzindo invariância espacial (KHAN *et al*., 2020).

No mesmo ano, os pesquisadores passaram a usar GPU's (processadores gráficos) para acelerar o treinamento de redes neurais profundas. Em 2007, a empresa NVIDIA lançou o CUDA: uma plataforma de programação que permite explorar as capacidades da computação paralela. As melhorias nas condições de hardware foi o principal fator para que as pesquisas em CNN's fossem retomadas (KHAN *et al*., 2020).

Em 2010, o grupo de Fei-Fei Li em Stanford criou um grande banco de dados de imagens, conhecido como ImageNet, contendo milhões de imagens rotuladas. Esse

banco de dados foi associado a uma Competição de Reconhecimento Visual de Grande Escala (ILSVRC), onde a performance de diversas redes é anualmente avaliada. Essa competição tem um importante papel pois além de fornecer um banco de dados para treinar as redes, tornou possível comparar tecnologias no estado da arte (KHAN *et al*., 2020).

3.4.4 A Ascenção das CNNs: 2012 – 2014

Em 2012, uma CNN chamada AlexNet venceu a competição ILSVRC, superando outras técnicas de *machine learning*. A arquitetura da AlexNet tinha o objetivo de melhorar o desempenho da rede ao explorar sua profundidade, incorporando vários níveis de transformações (KRIZHEVSKY; SUTSKEVER; HINTON, 2017). A performance exemplar dessa rede sugeriu aos pesquisadores que o principal motivo da saturação das CNNs em 2006 foi devido à falta de recursos computacionais e a quantidade insuficiente de dados necessários para um treinamento satisfatório, intrínsecos a época (KHAN *et al*., 2020).

Com as CNNs nos holofotes, uma série de tentativas de aprimoramento surgiram. Cada nova arquitetura emergia na tentativa de superar a anterior, combinando e/ou reformulando a sua estrutura. Em 2013 e 2014 os pesquisadores focaram principalmente na otimização de parâmetros, com pequeno aumento na complexidade computacional (KHAN *et al*., 2020).

Em 2014, um grupo de Oxford apresentou uma arquitetura chamada VGG. Em comparação com a AlexNet, a VGG utilizava filtros de tamanho muito menor. E o número de camadas passou de 9 para 16. Essa rede passou a lidar com um grande volume de mapas de ativação (SIMONYAN; ZISSERMAN, 2015). No mesmo ano, GoogleNet – vencedora do ILSVRC 2014 – foi além dos esforços em melhorias de design, essa rede introduziu os conceitos de transformação multinível: através de blocos baseados em divisão, transformação e mesclagem, essa técnica incorporou a rede a capacidade de capturar informações globais e locais, ajudando-a a lidar com detalhes da imagem em diferentes níveis (SZEGEDY *et al*., 2015; KHAN *et al*., 2020).

3.4.5 A descoberta das FCN: 2014

Em 2014, as pré-publicações de um trabalho revolucionário mudou a modo de se pensar as redes neurais de convolução. *Fully Convolutional Networks for semantic segmentation* foi oficialmente publicado em 2017. Apesar existirem trabalhos anteriores que usavam CNN's com o propósito de segmentação, pode-se dizer que a estrutura de rede proposta pelos autores desse artigo foi um divisor de águas (LONG; SHELHAMER; DARRELL, 2017).

Um ano antes, Lin *et al*. (2014) propôs a ideia de retirar as camadas totalmente conectadas das CNNs, com uma arrojada estrutura chamada *Network in Networks* (NIN), mas o objetivo ainda era a classificação.

Por outro lado, a ideia de Long, Shelhamer & Darrell (2017) era desmontar a camada totalmente conectada (do inglês *Fully Connected Layers*) das CNNs e substituí-las por uma nova camada de convolução, dando origem as chamadas Redes Totalmente Convolucionais (FCN). O objetivo era realizar a segmentação semântica de imagens fazendo uma adaptação na estrutura de redes neurais de classificação já conhecidas como AlexNet, VGG e GoogleNet. As arquiteturas mais amplamente utilizadas obtidas por este estudo foram derivadas da arquitetura VGG, conhecidas como "FCN-32s", "FCN-16s", "FCN-8s" (LONG; SHELHAMER; DARRELL, 2017).

3.4.6 Abordagens Pós-FCN: 2014 – Atualidade

Nos últimos 5 anos tem se observado um aumento drástico no interesse global pelo assunto de segmentação semântica. Quase todas as subsequentes abordagens seguem a ideia das Redes Totalmente Convolucionais (FCN). Não seria errado afirmar que, para o propósito de segmentação, as camadas totalmente conectadas (FC) deixaram de existir (ULKU; AKAGUNDUZ, 2020). Por outro lado, a ideia das FCNs também permitiu novas oportunidades para melhorar ainda mais as arquiteturas de segmentação, como é abordado no item 3.7.

3.5 A estrutura de rede básica para segmentação de imagens: FCN

As Redes Totalmente Convolucionais (do inglês FCN) surgiram a partir de uma mudança estrutural na composição das CNN's convencionais de classificação. Sua história é recente, e seu propósito é outro: ao invés de classificar a imagem como um todo, sua finalidade é segmentá-la entre diferentes classes (LONG; SHELHAMER; DARRELL, 2017).

Na tarefa de segmentação semântica, cada *pixel* da imagem é classificado, possuindo sua própria distribuição de probabilidades: a rede é treinada para prever a qual classe cada *pixel* pertence. Isso permite não apenas que uma variedade de classes seja identificada numa mesma imagem, mas também determinar sua localização (HEINRICH, 2016).

Uma imagem segmentada se assemelha a imagem original quanto a sua forma, e se difere quanto a cor. Na imagem segmentada, cada *pixel* é colorido com uma de *N* cores, onde *N* é o número de classes que se está segmentando. Ao segmentar uma imagem de estrada, por exemplo, isso pode ser tão simples como *N* = 2 ("estrada", ou "não-estrada"), como também pode capturar uma série de diferentes classes (figura 20).

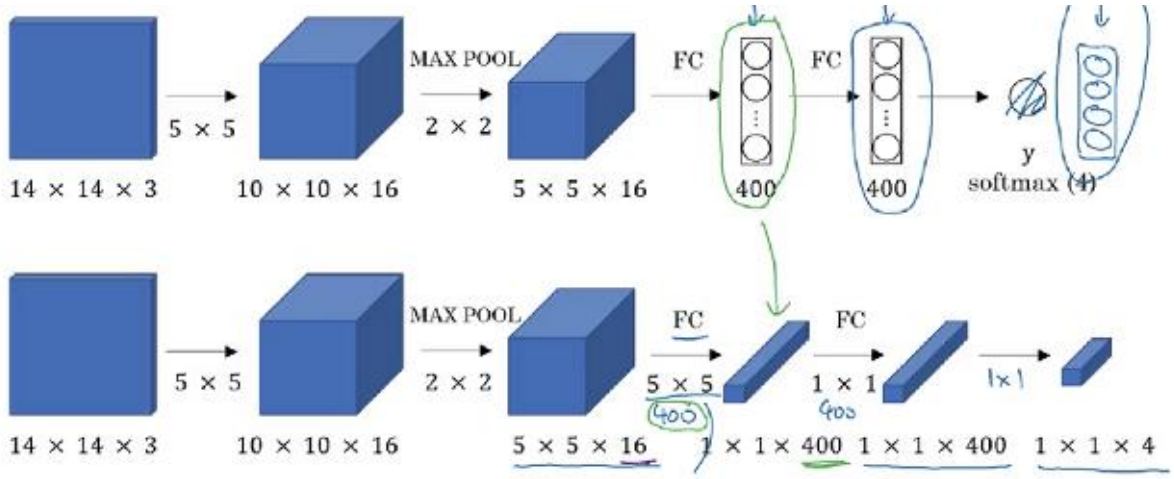
Figura 20 - Imagem segmentada usada no treinamento de redes neurais.

Extraído de: Site institucional do *The cityscapes dataset*.

É comum na literatura encontrar referência às redes típicas de classificação como CNN's, e às redes de segmentação como FCN's. Isso é uma aproximação, tendo em vista que existem muitos outros usos para CNN's e novos tipos de redes de segmentação.

Na realidade, alguns pequenos ajustes já são suficientes para tornar uma rede neural de classificação em uma rede neural de segmentação. A ideia chave para realizar essa transformação é analisar a estrutura de uma CNN de classificação e perceber que a camada totalmente conectada (FC) pode ser vista como uma camada de convolução, onde o tamanho dos filtros é igual ao tamanho dos mapas de ativação (figura 21) (LONG; SHELHAMER; DARRELL, 2017).

Figura 21 - Enxergando CNNs como FCNs.

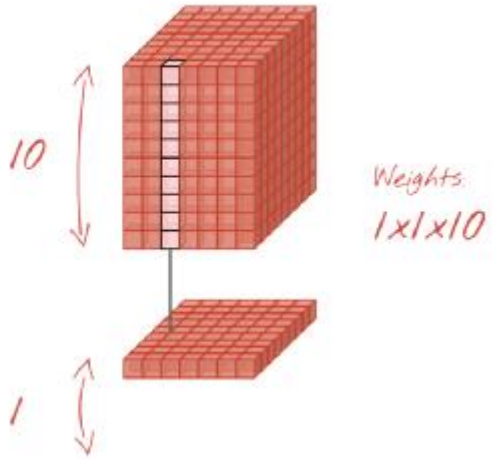


Extraído de: NG, 2017.

Nesse exemplo, retirado da apresentação de Andrew Ng (2017), ao realizar uma convolução com 400 filtros tamanho 5x5x16 em um bloco de dados de dimensões 5x5x16, é dado origem a um vetor de dados 1x1x400, fazendo com que o caráter espacial se perca. É o processo de achatamento ou *flattening* mencionado anteriormente.

No caso da segmentação de imagens, ao invés de analisar todas as características extraídas num vetor e retornar uma única divisão de probabilidades para a imagem, é necessário analisá-la *pixel* a *pixel*, e para isso é preciso substituir esses elementos que caracterizam a camada totalmente conectada. A nova camada de convolução utilizará filtros que ao invés de cobrir todo o bloco, cobrirão apenas 1 *pixel* ao longo do bloco. Pode-se dizer que será feita uma convolução unidimensional na profundidade do bloco (figura 22) (LONG; SHELHAMER; DARRELL, 2017; HEINRICH, 2019).

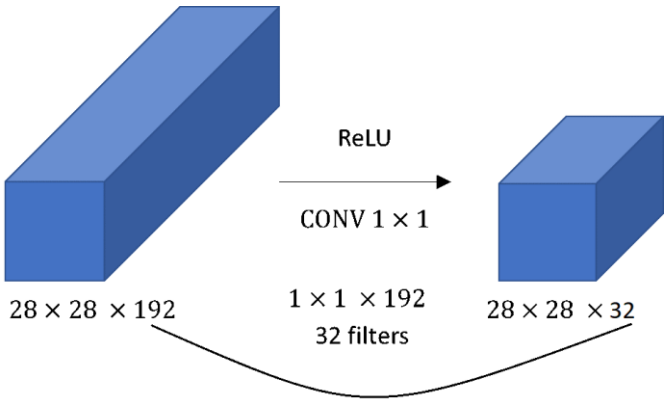
Figura 22 - Convolução unidimensional na profundidade do bloco.



Extraído de: GÖRNER, 2018

A nova camada será composta por *N* filtros de tamanho 1x1xC (onde *C* é o número de canais do bloco de dados). A ideia é analisar todas as ativações produzidas em cada *pixel* isoladamente, sendo o resultado um novo bloco de dados com número de canais igual a *N*. Se estamos classificando cada *pixel* como uma de *N* = 32 classes por exemplo, essa camada gera um bloco de dados com 32 canais (figura 23), e a classe mais provável para o *pixel* é aquela em que o respectivo canal obtém maior valor. Por fim, é realizado um *max pooling* unidimensional na profundidade desse bloco, dando origem a estrutura básica de segmentação de imagens via redes neurais profundas (LONG; SHELHAMER; DARRELL, 2017).

Figura 23 – Convolução com 32 filtros 1x1x192.



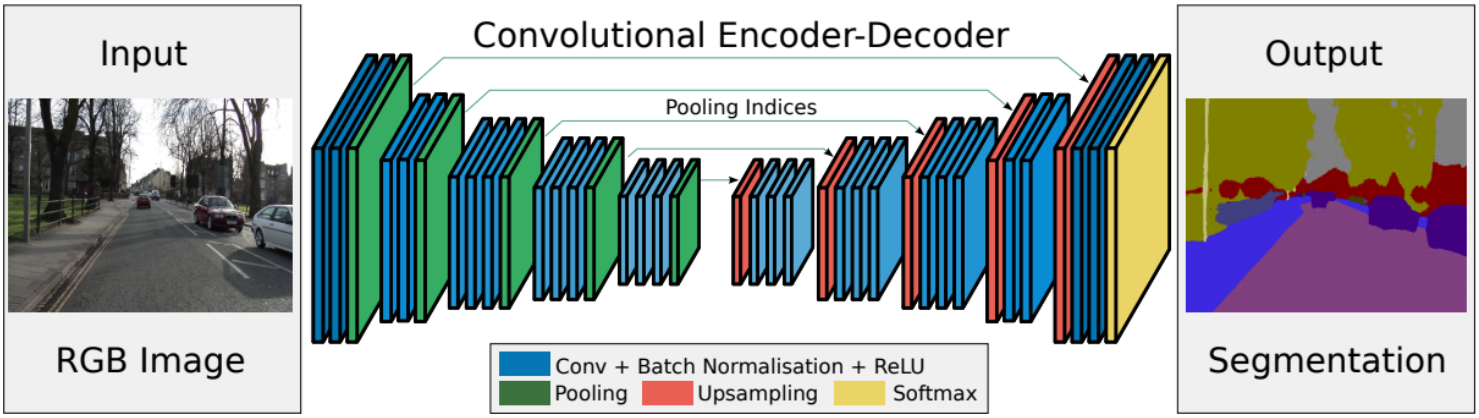
Extraído de: NG, 2017.

3.6 Abordagem pós-FCN: Redes SegNet

Na prática, para que o resultado de uma segmentação seja fidedigno, é vital que o delineamento das bordas do conteúdo da imagem segmentada esteja similar ao da imagem de entrada (BADRINARAYANAN; KENDALL; CIPOLLA, 2017). Para isso, as dimensões da imagem segmentada, devem ser iguais ou próximas às dimensões da imagem de entrada. Por outro lado, é sabido que os processos de convolução e *pooling* utilizados reduzem consideravelmente as dimensões dos blocos de dados. Isso nos conduz a um dos grandes desafios da segmentação de imagens: como aumentar a largura e altura da imagem segmentada a fim de atingir as dimensões originais da imagem de entrada, atentando-se à possíveis perdas na resolução quanto ao delineamento do conteúdo da imagem.

A essa tarefa se dá o nome de *upsampling*. Apesar das primeiras FCN's propostas por Long *et al*. (2017) já utilizarem métodos de *upsampling* através de convoluções transpostas, os estudos mais recentes no campo vêm mostrando diferentes abordagens para realizar essa tarefa, uma delas pode ser encontrada nas redes SegNet (BADRINARAYANAN; KENDALL; CIPOLLA, 2017).

Figura 24 - Arquitetura das redes SegNet.

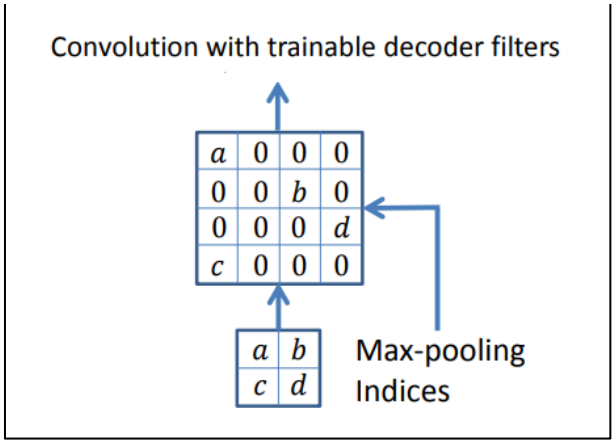


Extraído de: BADRINARAYANAN; KENDALL; CIPOLLA, 2017

A arquitetura SegNet possui uma estrutura chamada *encoder-decoder* (figura 24). Na primeira metade, a rede consiste em 13 camadas de convolução exatamente iguais às de uma rede de classificação já conhecida (VGG), entretanto, para cada mapa de ativação, os índices das localizações dos valores máximos computados durante o *max pooling* são armazenados (BADRINARAYANAN; KENDALL; CIPOLLA, 2017).

O *decoder* realiza o *upsampling* a partir dos índices armazenados (*unpooling*), e aplica convoluções após cada aumento: os filtros de convolução do *decoder* são treinados com o propósito de densificar os mapas esparsos criados com essa técnica (figura 25) (BADRINARAYANAN; KENDALL; CIPOLLA, 2017).

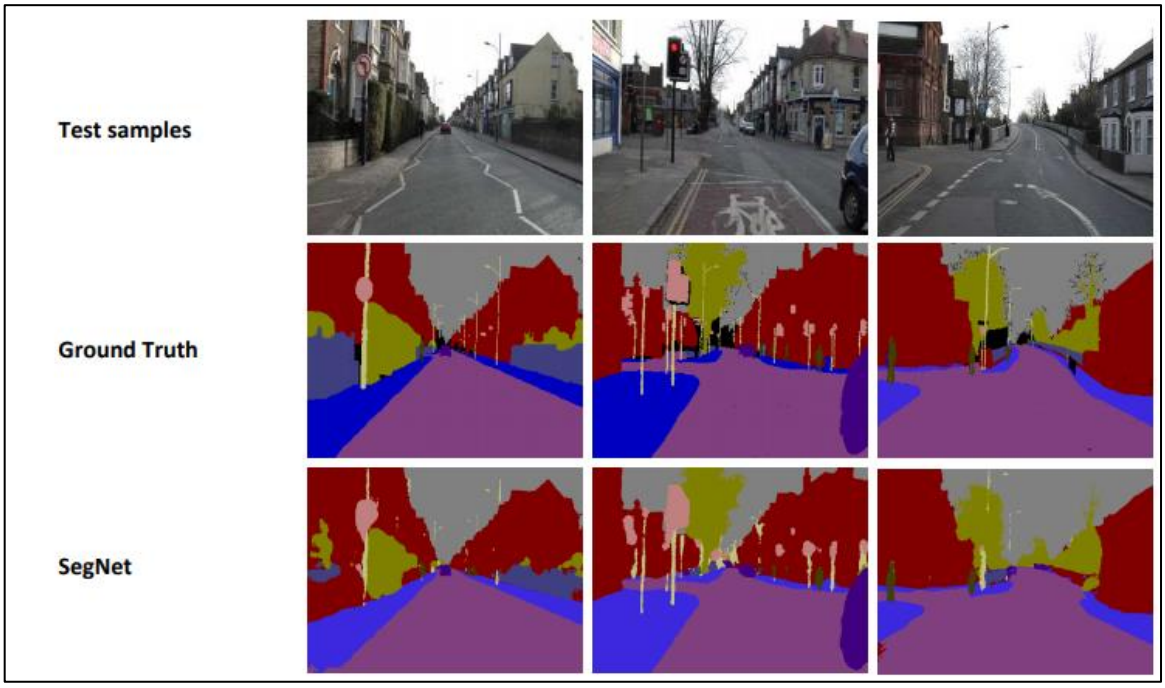
Figura 25 – Processo de *unpooling* realizado nas SegNet



Extraído de: BADRINARAYANAN; KENDALL; CIPOLLA, 2017

Na estrutura SegNet, o número de mapas de ativação no *encoder* e respectivo *decoder* são iguais, exceto no último *decoder*, onde são gerados vários mapas de ativação apesar da entrada no primeiro *encoder* possuir 3 canais (RGB). Essa variedade de mapas gerados alimenta um classificador softmax, que retorna uma imagem com N canais, onde N é o número de classes. A segmentação prevista corresponde às classes com maior probabilidade para cada *pixel*. Na figura 26 é possível visualizar o resultado de segmentação de algumas amostras de teste (BADRINARAYANAN; KENDALL; CIPOLLA, 2017).

Figura 26 – Amostras de teste, gabarito, e a saída da rede SegNet.



Extraído de: BADRINARAYANAN; KENDALL; CIPOLLA, 2017

Além da estrutura *encoder-decoder*, outras técnicas vêm sendo exploradas no campo de segmentação de imagens, cada uma com suas vantagens e desvantagens, que podem ser exploradas dependendo da aplicação. Alguma delas são: *Spatial Pyramid Pooling*, *Fine-grained localisation*, *Feature concatenation*, *Dilated convolution e Conditional Random Fields* (ULKU; AKAGUNDUZ, 2020).

3.7 O treinamento das redes e seu papel na evolução das CNNs

As redes neurais aprendem correlações entre dados através de treinamento: numa rede, para cada interação entre dois neurônios é atribuído um peso, um fator multiplicativo. De modo geral, o treinamento consiste em ajustar esses pesos, a fim de reduzir o erro observado ao comparar a saída da rede com a previsão esperada. Numa CNN, tanto os pesos nas camadas totalmente conectadas, quanto os filtros da convolução são estabelecidos via treinamento.

3.7.1 *Backpropagation* em CNNs

As redes neurais convolucionais foram uma das primeiras redes profundas funcionais treinadas com *back-propagation* (GOODFELLOW; BENGIO; COURVILLE, 2016). De modo que ainda em 1989, Yann LeCun *et al* (1989) já haviam constatado empiricamente o potencial do algoritmo no treinamento de CNNs, observando uma convergência rápida e tempo de treinamento razoáveis. Nessa época, as aplicações se restringiam a reconhecimento de dígitos manuscritos.

Em 2006, Hinton *et al.* (2006) propôs *greedy layer-wise pretraining*: um método de treinamento não supervisionado onde não se propagava os dados por toda a rede para cada passo do gradiente. Nessa abordagem, a primeira camada era treinada isoladamente, e então após extrair as características dessa primeira camada, a segunda camada era treinada isoladamente levando em consideração as características da primeira. Esse método ficou popular entre 2007 e 2013, época em que o poder computacional era notadamente limitado e os bancos de imagens para treinamento, muito pequenos. (GOODFELLOW; BENGIO; COURVILLE, 2016).

Hoje, a maioria das redes convolucionais fazem uso do mesmo algoritmo de treinamento supervisionado, utilizado a décadas atrás. A rede aprende através de propagações completas (*foward* e *backpropagation*) a cada iteração do treinamento (GOODFELLOW; BENGIO; COURVILLE, 2016).

A melhora da performance e evolução das CNNs desde 1989 a 2015 pode ser atribuída a dois fatores. Primeiro, maiores bancos de imagens diminuíram o grau de generalização das redes. Segundo, as arquiteturas de rede se tornaram muito maiores, devido a computadores mais poderosos e software com melhores infraestruturas (GOODFELLOW; BENGIO; COURVILLE, 2016).

3.7.2 *ImageSets* ou bancos de imagens

No treinamento de uma CNN, um conjunto de imagens de entrada com gabarito (anotadas) são utilizados. A qualidade e quantidade desses dados tem papel determinante no nível de sucesso da rede (ULKU; AKAGUNDUZ, 2020).

No caso das redes para segmentação de imagens, esse treinamento depende de imagens anotadas *pixel* a *pixel*. Classificar manualmente esse tipo de imagem é bastante demorado, frustrante e comercialmente caro (GUO *et al*., 2018). Daí a importância dos chamados *ImageSets*, ou banco de imagens.

Apesar dos vários campos de aplicação para problemas de segmentação, como imagens médicas, imagens de satélite e etc., o desenvolvimento científico no campo vem sendo guiado pelo setor de imagens de ruas urbanas, pois atraem mais a atenção da indústria e, portanto, uma série de banco de imagens e competições foram criadas para essa finalidade (ULKU; AKAGUNDUZ, 2020).

De acordo com Ulku e Akaganduz (2020), os bancos de imagens podem ser categorizados em dois ramos principais: banco de imagens com propósito geral, com classes genéricas que incluem quase todo tipo de objeto e plano de fundo; e banco de imagens de ruas urbanas, que incluem classes como carros e pedestres, e são geralmente usadas para treinar sistemas de carros autônomos (ULKU; AKAGUNDUZ, 2020).

3.7.2.1 Banco de imagens com propósito geral

• PASCAL *Visual Object Classes* (VOC): Esse banco de imagens inclui anotações não apenas para segmentação semântica, mas também para as tarefas de classificação, detecção, classificação de ação, e de partes do corpo. O banco de imagem e as anotações são regularmente atualizados e o *ranking* da competição é público (ocupado por mais de 100 ranqueados apenas para a tarefa de segmentação). É a mais popular entre as competições de segmentação semântica, e segue ativo desde seu lançamento em 2005. O banco de imagens da competição PASCAL VOC de segmentação semântica inclui 20 classes de objeto e uma classe de plano de fundo. O banco foi originalmente constituído por 1464 imagens de treinamento e outras 1449 imagens para validação. As 1456 imagens de teste são mantidas em segredo para a competição. O banco inclui todos os tipos de ambientes, fechados e abertos (EVERINGHAM *et al*., 2010; ULKU; AKAGUNDUZ, 2020).

• *Common Objects in Context* (COCO): Com 200 mil imagens anotadas, 1,5 milhões de instâncias de objeto e 80 categorias de objeto, COCO é um banco de imagens de larga escala, e abrange tarefas de segmentação semântica e detecção de objetos. Inclui quase todos os tipos possíveis de cena. O *ranking* das competições é relativamente vazio devido à grande quantidade de dados e classes, sendo ocupado apenas pelos cientistas mais ambiciosos (LIN *et al*., 2014; ULKU; AKAGUNDUZ, 2020).

• *Open Images Dataset V6* +: Com aproximadamente 9 milhões de imagens anotadas, esse banco de imagens inclui anotações de instâncias de objeto e máscaras de segmentação, além de novos tipos de anotações para diferentes tarefas como relacionamentos visuais e narrativas localizadas. As imagens são muito variadas e contém descrições mais específicas das imagens quando comparada aos demais bancos. As fronteiras dos objetos anotados são demarcadas com precisão, esse detalhamento é importante pois favorece o aprendizado dos modelos.

Apesar de menos populares que o PASCAL VOC, COCO, ou *Open Images V6,* existem outros bancos de dados de propósito geral para segmentação semântica.

YouTube-Objects é um deles: um conjunto de videoclipes em baixa resolução (480 x 360) com mais de 10 mil frames anotados *pixel* a *pixel*. Outro banco com imagens em baixa resolução (256 x 256) chama-se *SIFT-flow*, possuindo 2688 imagens para 33 diferentes classes de objeto. Além desses, outros bancos relativamente primitivos foram praticamente abandonados na literatura de segmentação semântica devido as baixas resoluções e volume (LAZEBNIK, 2010; PREST *et al*., 2012; TIGHE; ULKU; AKAGUNDUZ, 2020).

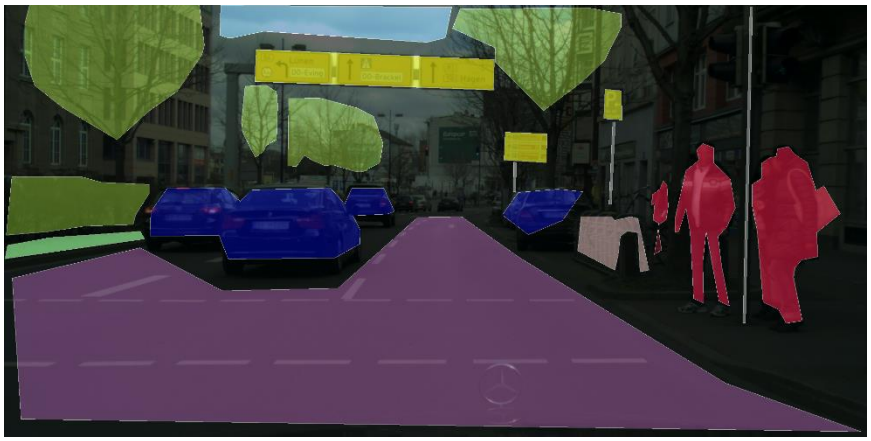
3.7.2.2 Banco de imagens de ruas urbanas

• *Cityscapes*: Esse banco de imagens de larga escala com foco no entendimento do ambiente urbano contém anotações em imagens de alta resolução, tiradas em 50 cidades, em diferentes horários do dia e em todas as estações do ano, com variações no plano de fundo. As anotações são divididas em duas categorias: “boas” (figura 27) para 5 mil imagens, e “grosseiras” (figura 28) para 20 mil. Existem 30 classes diferentes, algumas com anotações de diferentes instâncias. Consequentemente existem duas competições com *rankings* separados: uma para segmentação semântica e outra para segmentação semântica com identificação de instâncias. Com mais de 100 redes ranqueadas, esse é o banco de imagens mais popular para segmentação semântica de ruas urbanas (CORDTS *et al*., 2016; ULKU; AKAGUNDUZ, 2020).

Figura 27 – Uma amostra de anotação “boa” do *Cityscape ImageSet* (“*fine annotation*”).

Extraído de: Site institucional do *The cityscape dataset*.

Figura 28 - Uma amostra de anotação “grosseira” do *The cityscape ImageSet* (“*coarse annotation*”).



Extraído de: Site institucional do *The cityscape dataset*.

Outros bancos de imagens para ruas urbanas como *CamVid*, KITTI e SYNTHIA foram ofuscados pelo *Cityscapes* por diversas razões, principalmente devido as baixas resoluções. Desses bancos, apenas o SYNTHIA pode ser considerado de larga escala (com mais de 13 mil imagens anotadas), entretanto, são imagens geradas artificialmente e isso é considerado uma limitação para aplicações que exigem um nível de segurança crítico como em carros autônomos (BROSTOW; FAUQUEUR; CIPOLLA, 2009; GEIGER *et al*., 2013; ROS *et al*., 2016; ULKU; AKAGUNDUZ, 2020).

4 CONCLUSÃO

As Redes Neurais Convolucionais desempenharam um importante papel na história do *Deep Learning*. Foram uma das primeiras redes neurais profundas com aplicação comercial, tornando-se um exemplo bem-sucedido de solução inspirada na biologia.

Com o passar dos anos, a ação conjunta de mentes brilhantes e a metodologia científica permitiram que as arquiteturas de rede acompanhassem o avanço vertiginoso da tecnologia.

A pesquisa bibliográfica realizada para desenvolver o trabalho permitiu investigar as contribuições dos conhecimentos através de novas arquiteturas que emergiam. E tornou possível compreender as funções das diferentes camadas da rede:

• As camadas de convolução são responsáveis por extrair característica de baixo nível da imagem (como retas, pontas, curvaturas etc.). Através de filtros que percorrem a imagem.

• As camadas não-lineares favorecem o aprendizado de relações complexas entre as características extraídas.

• As camadas de *pooling* comprimem essas informações e garantem uma invariância espacial das características.

• Outras camadas de convolução extraem características cada vez mais abstratas ao combinar características anteriores.

• As camadas totalmente conectadas analisam as características mais abstratas extraídas pelas camadas anteriores, e retornam uma classificação para a imagem.

• Substituir a camada totalmente conectada por uma camada de convolução 1x1 possibilita a criação de uma rede capaz de segmentar imagens no estado da arte.

• Novas estruturas como as de decodificação permitem refinar a posição dos *pixels* segmentados pela rede neural profunda.

Por fim, foi entendida a importância dos diferentes bancos de imagens que existem hoje, essenciais no treinamento e, portanto, na criação de redes de convolução cada vez mais poderosas.

REFERÊNCIAS

HEBB, Donald. **The Organization of Behavior: A Neuropsychological Theory**. Nova Iorque: Wiley, 1949.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. v. 4. Ed. São Paulo: Atlas. 2008.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. MIT Press. 2016.

DESHPANDE, Adit. A Beginer's Guide to Understanding Convolutional Neural Networks. **Adit Deshpande Blog**. 2016. Disponível em: https://adeshpande3.github.io/adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks/. Acesso em: 20 jun. 2019.

HEINRICH, Greg. Image Segmentation Using DIGITS 5. **NVIDIA Developer Blog**. 2016. Disponível em: https://developer.nvidia.com/blog/image-segmentation-using-digits-5/. Acesso em: 12 ago. 2019.

GUO, Yanming *et al*. A Review of Semantic Segmentation Using Deep Neural Networks. **International Journal of Multimedia Information Retrieval**. v. 7, p. 87-93. 2018.

SZEGEDY, Christian *et al*. Going Deeper with Convolutions. **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Boston. p.1-9. 2015.

ULKU, Irem; AKAGUNDUZ, Erdem. **A Survey on Deep Learning-Based Architectures for Semantic Segmentation on 2D Images**. Pré-publicação. Çankaya University, Turquia, 2020.

BADRINARAYANAN, Vijay; KENDALL, Alex; CIPOLLA, Roberto. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. v. 39. p. 2481-2495. 2017.

KHAN, Asifullah *et al*. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. **Artificial Intelligence Review**. v. 53. p. 5455-5516. 2020

ARNAB, Anurag *et al*. Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation: Combining Probabilistic Graphical Models with Deep Learning for Structured Prediction. **IEEE Signal Processing Magazine**. v. 35. n. 1 p. 37-52, jan. 2018.

KUFFLER, Stephen. Discharge Patterns and Functional Organization of Mammalian Retina. **Journal of Neurophysiology**. v.16. n. 1. p. 37-68. jan. 1953.

LECUN, Yann *et al*. Backpropagation Applied to Handwritten Zip Code Recognition. **Neural Computation**. v. 1. n. 4. p. 541-551. 1989.

LECUN, Yann *et al*. Gradient-based Learning Applied to Document Recognition. **Proceedings of the IEEE**. v. 86. n. 11. p. 2278-2324. nov. 1998.

HINTON, Geoffrey; OSINDERO, Simon; TEH, Yee-Whye. A fast learning algorithm for deep belief nets. **Neural Computation**. v. 18. n. 7. p. 1527-1554. jun. 2006.

HUBEL, David; WIESEL, Torsten. Receptive fields of single neurones in the cat's striate córtex. **The Journal of Physiology**. v. 148. n. 3. p. 574-591. out. 1959.

HUBEL, David; WIESEL, Torsten. Receptive fields, binocular interaction and functional architecture in the cat's visual córtex. **The Journal of Physiology**. v. 160. n. 1. p. 106-154. jan. 1962.

SIMARD, Patrice; STEINKRAUS, Dave; PLATT, John. Best practices for convolutional neural networks applied to visual document analysis. **Seventh International Conference on Document Analysis and Recognition**. Edimburgo. p. 958-963. 2003.

RANZATO, Marc'Aurelio *et al*. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. **2007 IEEE Conference on Computer Vision and Pattern Recognition**. Mineápolis. p. 1-8. 2007.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey. ImageNet classification with deep convolutional neural networks. **Communications of the ACM**. v. 60. n. 6. maio 2017.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition**. 3rd International Conference on Learning Representations (ICLR)**. San Diego. 2015

LONG, Jonathan; SHELHAMER, Evan; DARRELL, Trevor. Fully Convolutional Networks for semantic segmentation. **2015 IEEE Conference of Vision and Pattern Recognition (CVPR)**. Boston. p. 3431-3440. 2015.

RIZWAN, Muhammad. Convolutional Neural Networks – In a Nut Shell. **engMRK**. 17 set. 2018. Disponível em: https://engmrk.com/convolutional-neural-network-3/ Acesso em: 5 fev. 2020.

JORDAN, Jeremy. Convolutional Neural Networks. **Jeremy Jordan**. 16 jul. 2017. Disponível em: https://www.jeremyjordan.me/convolutional-neural-networks/ Acesso em: 10 fev. 2020.

NG, Andrew. **Object Detection**. 27 nov. 2017. Apresentação de slides. Disponível em: https://x-wei.github.io/notes/Ng_DLMooc_c4wk3.html Acesso em: 16 jun. 2020.

EVERINGHAM, Mark *et al*. The pascal visual object classes (voc) challenge. **International Journal of Computer Vision**. v. 88. p 303-338. 2010.

LIN, Tsung-Yi *et al*. Microsoft COCO: Common objects in context**. European Conference on Computer Vision - ECCV 2014**. p.740-755. 2014.

PREST, Alessandro *et al*. Learning object class detectors from weakly annotated vídeo*.* **2012 IEEE Conference on Computer Vision and Pattern Recognition**. Providence. p. 3282-3289. 2012.

TIGHE, Joseph; LAZEBNIK, Svetlana. Superparsing: Scalable nonparametric image parsing with superpixels. **European Conference on Computer Vision – ECCV 2010**. p. 352-365. 2010.

CORDTS, Marius *et al*. The cityscapes dataset for semantic urban scene understanding. **Proceedings of the IEEE conference on computer vision and pattern recognition**. p. 3213-3223. 2016.

BROSTOW, Gabriel; FAUQUEUR, Julien; CIPOLLA, Roberto. Semantic object classes in video: A high-definition ground truth database. **Pattern Recognition Letters**. v. 30. p. 88-97. 2009.

GEIGER, Andreas *et al*. Vision meets robotics: The KITTI dataset. **The International Journal of Robotics Research**. v. 32. n. 11. p. 1231-1237. ago. 2013.

ROS, German *et al*. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Las Vegas. p. 3234-3243. 2016.

ÜNAL, Mehmet. Show Images Directly on Terminal: img2sh. **A blog from Mehmet Ozan Ünal**. 3 nov. 2019. Disponível em: https://mozanunal.com/2019/11/img2sh/ Acesso em: 5 set. 2020.

GÖRNER, Martin. **Modern Convolutional Neural Nets**. 2018. Apresentação de slides. Disponível em: https://github.com/GoogleCloudPlatform/tensorflow-without-a-phd Acesso em: 10 set. 2020.

ANIEMEKA, Ifu. A Friendly introduction to Convolutional Neural Networks. **Hashrocket blog**. 22 ago. 2017. Disponível em: https://hashrocket.com/blog/posts/a-friendly-introduction-to-convolutional-neural-networks Acesso em: 15 ago. 2020.

THE CITYSCAPES DATASET. [Site institucional]. Disponível em: www.cityscapes-dataset.com. Acesso em: 5 set. 2020.

LIN, Min; CHEN, Qiang; YAN, Shuicheng. Network in Network. **Computer Science**. 2014.

ZEILER, Matthew; FERGUS, Rob. Visualizing and Understanding Convolutional Networks. **European Conference on Computer Vision 2014.** p. 818-833. 2014