PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Informática

Victor Hugo Roldão Reis

PROPOSTA E ANÁLISE DE ARQUITETURAS DE REDES NEURAIS PROFUNDAS PARA A CONTAGEM DE MULTIDÕES

Belo Horizonte

2020

Victor Hugo Roldão Reis

PROPOSTA E ANÁLISE DE ARQUITETURAS DE REDES NEURAIS PROFUNDAS PARA A CONTAGEM DE MULTIDÕES

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Orientador: Prof. Dr. Zenilton Kleber Gonçalves do Patrocínio Júnior

Belo Horizonte

2020

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Reis, Victor Hugo Roldão
R375p Proposta e análise de arquiteturas de redes neurais profundas para a contagem de multidões / Victor Hugo Roldão Reis. Belo Horizonte, 2020.
92 f. : il.

Orientador: Zenilton Kleber Gonçalves do Patrocínio Júnior
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Computação evolutiva. 2. Processamento de dados. 3. Sistemas de computação. 4. Redes neurais (Computação). 5. Arquitetura de rede de computador. 6. Banco de dados. I. Patrocínio Júnior, Zenilton Kleber Gonçalves do. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 681.3.091

Ficha catalográfica elaborada por Pollyanna Iara Miranda Lima - CRB 6/3320

Victor Hugo Roldão Reis

PROPOSTA E ANÁLISE DE ARQUITETURAS DE REDES NEURAIS PROFUNDAS PARA A CONTAGEM DE MULTIDÕES

Dissertação apresentada ao Programa de Pós-Graduação em Informática como requisito parcial para qualificação ao Grau de Mestre em Informática pela Pontifícia Universidade Católica de Minas Gerais.

Prof. Dr. Zenilton Kleber Gonçalvez do Patrocínio Júnior – PUC Minas
(Orientador)

Prof. Dr. Silvio Jamil Ferzoli Guimarães – PUC Minas (Banca Examinadora)

Prof. Dr. Alexei Manso Correa Machado – PPGEE/PUC Minas (Banca Examinadora)

Prof. Dr. Sandra Eliza Fontes de Avila – Unicamp (Banca Examinadora)

Belo Horizonte, 13 de novembro de 2020.

Para aqueles que dedicaram e dedicam suas vidas a ensinar e a aprender.

AGRADECIMENTOS

Agradeço a todos os professores com quem tive a oportunidade de aprender, deste a professora América na minha infância até os mestres e doutores com quem pude conviver na graduação e pós-graduação como Paulo Amaral, Silvio Jamil, dentre vários outros. À minha família e amigos, principalmente ao amigo Cristiano Lacerda com quem pude muito aprender e me serviu de inspiração na vida. A todos os funcionários do Programa de Pós-Graduacao em Informática. Um agradecimento especial ao meu professor e orientador Zenilton Kleber Gonçalves do Patrocínio Júnior, pessoa por quem eu preservo enorme admiração e consideração. Sem ele, este trabalho não seria possível.

"Estamos tão familiarizados com a visão, que é preciso um salto de imaginação para perceber que existem problemas a serem resolvidos. Mas considere isto. Recebemos imagens minúsculas e distorcidas de cabeça para baixo nos olhos e vemos objetos sólidos separados no espaço circundante. A partir dos padrões de estimulação na retina percebemos um mundo de objetos, e isso não é nada menos que um milagre."

Richard L. Gregory

RESUMO

Um dos desafios atuais da área de visão computacional diz respeito a contagem de indivíduos em uma cena ou, como é amplamente conhecida, a contagem de multidões. Para resolver este problema computacionalmente, precisa-se de modelos capazes de estimar o número de indivíduos presentes na multidão representada por imagens digitais ou vídeos. Nestes cenários, quanto maior a aglomeração maiores serão os desafios na identificação dos indivíduos de forma isolada. Diferentes abordagens foram utilizadas na tentativa de solucionar este problema, destacando-se as baseadas em estimativa de densidade como aquelas que apresentam os melhores resultados, principalmente em cenários de alta densidade. Em geral, as redes neurais artificiais convolucionais tem sido amplamente utilizadas para estimar a densidade de multidões. No entanto, este tipo de rede apresenta algumas relevantes limitações como, por exemplo, a perda da localização das características ao longo da rede. Para minimizar esta perda, algumas arquiteturas sugerem a combinação de camadas em diferentes posições na arquitetura da rede. Além disso, como alternativa surgiram as redes em cápsulas na tentativa de melhor modelar as relações hierárquicas existentes em uma imagem e preservar informações da localização das características ao longo da rede. Modelos baseados em redes em cápsulas possuem forte capacidade de representação e um poderoso mecanismo de roteamento dinâmico que pode resolver a desvantagem de um número limitado de amostras de treinamento, que é um dos problemas que se enfrenta na tarefa de contagem de multidões. Neste trabalho, foram propostos três modelos de aprendizado supervisionado para a solução do problema da contagem de multidões: (i) baseado em redes em cápsulas; (ii) baseado em rede convolucional com arquitetura semelhante ao formato em "U" *(Quasi U-NET)*; e (iii) um modelo híbrido que faz a junção de redes neurais convolucionais que foram previamente treinadas, explorando a transferência de aprendizado, com uma rede inspirada na ideia de redes em cápsulas. Todas estas abordagens têm a mesma finalidade que é estimativa da contagem por meio de mapas de densidade, que poderão ser integrados para gerar o número aproximado de indivíduos na multidão. A abordagem híbrida se mostrou eficiente e foi capaz de alcançar o estado da arte em contagem de multidões e as outras abordagens ficaram próximas do feito.

Palavras-chave: Aprendizado Supervisionado. Contagem de Multidões. Redes Neurais Convolucionais. Mapas de Densidade. Redes em Cápsula.

ABSTRACT

One of the current challenges of computer vision concerns the counting of individuals in a scene or, as it is widely known, the counting of crowds. To solve this problem computationally, we need models capable of estimating the number of individuals present in the crowd represented by digital images or videos. In these scenarios, the greater the agglomeration, the greater the challenges in identifying individuals in isolation. Different approaches were used in an attempt to solve this problem, highlighting those based on density estimation as those that present the best results, especially in high-density scenarios. In general, convolutional neural networks have been widely used to estimate crowd density. However, this type of network has relevant limitations, such as the loss of the location of characteristics along the network. To minimize this loss, some architectures suggest the combination of layers in different positions in the network architecture. In addition, as an alternative, networks in capsules emerged in an attempt to better model the existing hierarchical relationships in an image and preserve information on the location of characteristics along the network. Models based on capsule networks have a strong representation capacity and a powerful dynamic routing mechanism that can solve the disadvantage of a limited number of training samples, which is one of the problems faced in the crowd counting task. In this work, three supervised learning models were proposed to solve the crowd counting problem: (i) based on capsule networks; (ii) based on a convolutional network with architecture similar to the "U" format *(Quasi U-NET)*; and (iii) a hybrid model that joins convolutional neural networks that were previously trained, exploring the transfer of learning, with a network inspired by the idea of networks in capsules. All of these approaches have the same purpose as estimating the count using density maps, which can be integrated to generate the approximate number of individuals in the crowd. The hybrid approach demonstrate to be efficient and was able to reach the state of the art in crowd counting and the other approaches were close to the feat.

Keywords: Supervised Learning. Crowd Couting. Convolutional Neural Network. Density Maps. Capsule Neural Network.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Ca-CCNet – *CapsNet for Crowd Counting Network*

CapsNet – *Capsule Neural Network*

CaTL-CCNet – *CapsNet with Transfer Learning for Crowd Counting Network*

CNN – *Convolutional Neural Network*

CSRNet – *Congested Scene Recognition Network*

EM – *Expectation-Maximization*

FLOP – *FLoating-point Operations Per Second*

GPU – *Graphic Processing Unit*

MAE – *Mean Absolute Error*

MLP – *Multilayer Perceptron*

MSE – *Mean Squared Error*

QU-CCNet – *Quasi U-Net for Crowd Counting Network*

TL – *Transfer Learning*

SUMÁRIO

1 INTRODUÇÃO

Sistemas de vigilância carecem de mecanismos eficientes para monitoramento da população de forma que lhes garanta segurança em lugares como shoppings, estádios, arenas, shows ou outros espaços públicos. Estes sistemas não só dependem de equipamentos (câmeras digitais) com boa resolução, mas também de técnicas capazes de alcançar resultados satisfatórios com os dados obtidos por estes equipamentos. Fatores como posição da câmera, iluminação e distância afetam diretamente a qualidade destes dados.

Uma das formas de monitoramento envolve a contagem dos indivíduos, pois através deste levantamento é possível definir estratégias para o controle, segurança, mobilidade pública e planejamento urbano. Em determinadas multidões pode haver comportamentos que mereçam mais atenção e o monitoramento permite a definição de estratégias para a proteção dos indivíduos contra ataques à sua integridade.

A visão computacional, no início dos anos 70, era vista como o componente de percepção visual de uma agenda ambiciosa para imitar a inteligência humana e dotar os robôs de comportamento inteligente. Alguns dos pioneiros acreditavam que resolver o problema da "entrada visual" seria um passo fácil para solucionar problemas mais difíceis, como planejamento e raciocínio de nível superior.

O crescimento do poder computacional e a ascensão da inteligência artificial contribuíram para o surgimento das redes neurais artificiais profundas que são algoritmos que simulam o mecanismo de aprendizado em organismos biológicos, muito utilizados no aprendizado de máquina. Nos últimos anos, modelos inspirados em rede neural tem sido amplamente utilizados e alcançaram bons resultados em tarefas que envolvem o processamento e análise de imagens digitais.

A contagem de multidões é um dos desafios da área de visão computacional que tem aplicações não somente nas áreas de segurança e planejamento, mas também em outras áreas do conhecimento que envolvem aspectos sociais e psicológicos relacionados ao comportamento das pessoas contidas na multidão analisada.

A maneira que os indivíduos se caracterizam na cena afeta diretamente a escolha do método computacional a ser utilizado para resolver o problema da contagem de multidões. Estas características dizem respeito, basicamente, à quantidade de indivíduos na cena, a posição e distância que estão da câmera. Alguns exemplos destas características, presentes nas imagens utilizadas neste tipo de tarefa, estão em destaque na Figura 1.

Figura 1 – Exemplo de cenas de multidões em diferentes pontos de vista, extraídas das bases de imagens utilizadas para este tipo de tarefa.

Fonte: Elaborada pelo autor

Outras características, importantes no trabalho de contagem de multidões, dizem respeito ao tamanho da cabeça do indivíduo na cena, a distância de cada um dos indivíduos e a oclusão de parte da sua cabeça.

Para Loy et al. (2013), os métodos tradicionais de contagem de multidões podem ser classificados em: métodos baseados em detecção, métodos baseados em regressão e métodos baseados em estimativa de densidade.

Os métodos baseados em detecção em geral são utilizados em cenas que possibilitam a identificação isolada dos indivíduos. No entanto, estes métodos falham na medida em que a multidão se torna mais densa.

Segundo Li, Zhang e Chen (2018), ao executar uma solução baseada em regressão, uma característica muito importante denominada saliência é negligenciado, o que causa resultados imprecisos em determinadas regiões.

Em geral, os mapas de densidade são uma alternativa eficiente em cenários na qual os indivíduos não são facilmente identificáveis, principalmente quando a multidão é mais densa. Lempitsky e Zisserman (2010) sugerem o mapeamento linear entre características locais e seus mapas de densidade integrando as informações de saliência durante o processo de aprendizado. Portanto, a correta estimativa de densidade passa a ser uma questão chave no desenvolvimento de um método para contagem de multidões.

Das abordagens tradicionais, aquelas que utilizam redes neurais convolucionais, do inglês *Convolutional Neural Network* (CNN), têm alcançados os melhores resultados. Elas representam uma abordagem de aprendizado profundo na qual uma arquitetura de rede neural busca aferir o número de indivíduos na cena.

Para a solução deste problema, foram propostas diferentes abordagens. Existem não somente modelos construídos com base em arquiteturas novas treinadas a partir do zero, como as propostas por Zhang et al. (2016) e Liu et al. (2018), mas também modelos baseados em adaptações de redes já conhecidas e treinadas para resolver outros problemas por meio da transferência de aprendizado, do inglês *Transfer Learning* (TL), como as propostas por Li, Zhang e Chen (2018), Valloli e Mehta (2019) e Shi et al. (2018).

Esta capacidade de transferência de aprendizado permite que um conhecimento prévio (características, pesos, etc.), alcançado por treinamento, validação e testes realizados anteriormente na solução de outras tarefas, possa ser utilizado para resolver outros problemas com, supostamente, um menor esforço principalmente na etapa de treinamento.

Um dos pontos que motivaram críticas nas CNNs se referem à perda de informações da localização das características (*features*) ao longo da rede devido, principalmente, a excessivas camadas de *pooling*, que reduzem a dimensão em um valor escalar que corresponde, por exemplo, ao valor máximo (*max-pooling*) ou médio (*average-pooling*) de determinada região de acordo com uma janela (*kernel*) previamente configurada.

Na tentativa de reduzir esta perda, algumas estratégias foram propostas. Uma delas envolve a construção de redes neurais artificiais profundas que combinam em sua arquitetura informações de camadas localizadas em diferentes níveis da rede, buscando melhorar a qualidade dos resultados. Outras estratégias se revelam necessárias na medida que esta combinação de camadas por si só não é o suficiente, principalmente em tarefas mais desafiadoras.

Como alternativa as CNNs tradicionais surgiram as redes em cápsulas, do inglês *Capsule Neural Network* (CapsNet), que são um tipo de rede neural artificial profunda inspiradas no conceito de cápsulas para melhor modelar as relações hierárquicas existentes em uma imagem analisada. De acordo com Sabour, Frosst e Hinton (2017), esta proposta sugere uma forma de roteamento mais efetiva do que a "forma primitiva" de roteamento alcançada por uma camada de *pooling*. Além disso, este tipo de rede supostamente precisaria de menos dados para aprender a representação dos objetos.

Sendo assim, neste trabalho foram propostas três arquiteturas de redes neurais artificiais profundas para a solução do problema da contagem de multidões: (i) baseada em CapsNet para a construção mapas de densidade de maior qualidade com uma abordagem que tenta modelar de maneira mais efetiva a relação hierárquica; (ii) com uma CNN que tenta combinar em sua arquitetura informações de diferentes camadas da rede; e (iii) uma rede que combina o poder da transferência de aprendizado de uma CNN previamente treinada com outra parte baseada em CapsNet.

Desenvolver modelos capazes de solucionar o problema da contagem de multidões, levando em consideração todos os desafios apresentados, de maneira computacionalmente

eficiente, é o desafio que se propõe superar nesta dissertação. Portanto, o conteúdo deste trabalho está focado no desenvolvimento e análise de arquiteturas de redes neurais artificiais profundas para a solução do problema da estimativa de contagem por mapas de densidade.

Neste Capítulo serão apresentadas a definição do problema (ver Seção 1.1), a motivação para a realização dessa pesquisa (ver Seção 1.2), os objetivos (ver Seção 1.3), a justificativa do trabalho (ver Seção 1.4) e as contribuições (ver Seção 1.5). Por fim, a organização da dissertação é apresentada na Seção 1.6.

1.1 Problema

Uma multidão pode se apresentar de diferentes maneiras a cada cena analisada. Ela pode ser vista de uma forma macro, como uma entidade que se modifica a cada instante devido ao comportamento dos diversos e diferentes indivíduos que a constituem.

Se analisada de diferentes maneiras, uma mesma cena de multidão pode também produzir diferentes resultados no que diz respeito a contagem. Em um dado instante pode se apresentar densa, ou seja, cheia de indivíduos, em seguida se tornar esparsa pela dispersão dos mesmos. Um exemplo deste problema está exposto na Figura 2.

Esta característica da multidão, isto é a forma como ela se apresenta, interfere diretamente na abordagem adotada para a solução do problema da estimativa de contagem.

Figura 2 – Exemplo de cenas de multidões representadas em diferentes configurações de perspectiva, pontos de visão e também da quantidade de indivíduos.

Fonte: Elaborada pelo autor

Os indivíduos presentes na multidão podem ser de diferentes etnias, tamanhos e cores (devido à iluminação do ambiente, cor de cabelos, de chapéus, etc.) conforme destaca a Figura 3. Podem ser vistos de frente, de lado, de costas, de cima, enfim... de diferentes maneiras. Suas expressões podem ser diversas (sorrindo, chorando, gritando, caladas, sérias, etc) ou nem sempre perceptíveis.

Em muitos casos, conseguimos observar somente partes do corpo, por isso as anotações contidas nas bases de imagem não são totalmente precisas no sentido de identificação de todos os indivíduos que, supostamente, estariam na cena. As anotações também podem variar sensivelmente em relação as coordenadas da posição na cabeça do indivíduo, ou seja, algumas marcações estão mais próximas dos olhos, outras da testa, outras mais próximas da boca, e isto pode de certa forma impactar nos resultados.

Um método proposto deve ser capaz também de distinguir indivíduos que possuem marcações realizadas na nuca ou na parte traseira da cabeça, de regiões que correspondem ao plano de fundo, devido a semelhança na textura em ambos os casos. Um exemplo deste problema está exposto na Figura 4.

Existem diferentes abordagens para lidar com o problema de contagem de multidões. Os métodos baseados em detecção são utilizados em cenas de multidões esparsas, no qual os indivíduos não se apresentam longe da câmera e podem ser facilmente identificados em seu todo e partes. No entanto, quando o número de indivíduos aumenta na multidão, este tipo de abordagem se revela ineficiente pela dificuldade em se identificar um indivíduo isoladamente. Podemos perceber este problema na Figura 5.

Em cenários de média e alta densidade, onde os indivíduos não são facilmente identificáveis, os mapas de densidade tem alcançado melhor desempenho na estimativa

Figura 3 – Cenas de multidões que contém pessoas de diferentes raças e etnias, com diferentes expressões faciais, quando possível de identificar. Em alguns casos a imagem está representada em escala de cinza, o que torna difícil a diferenciação do tom de pele.

Fonte: Elaborada pelo autor

Figura 4 – Uma cena de multidão na qual os indivíduos podem ser identificados mas não foram marcados (a); e outra cena que contém indivíduos anotados na região da nuca ou parte traseira da cabeça(b).

Fonte: Elaborada pelo autor

de contagem. Estes mapas são construídos inicialmente de acordo com as anotações das posições das cabeças dos indivíduos na cena.

Em geral, as bases de dados utilizadas neste tipo de tarefa fornecem este dado. Um algoritmo então aprende a partir destes mapas iniciais a produzir mapas de densidade para as novas imagens de multidões sem a necessidade das anotações utilizadas no treinamento.

No entanto, existe uma dependência sobre o domínio dos dados que pode interferir na escolha dos projetos arquitetônicos desenvolvidos para a solução do problema de contagem, ou seja, a limitação das imagens que constituem as bases de treinamento, exigem novas estratégias. Uma delas envolve a construção de modelos capazes de lidar

Figura 5 – Cenas de multidões densas em que os métodos por detecção falham na medida em que não conseguem identificar os indivíduos isoladamente.

Fonte: Elaborada pelo autor

com estas limitações. Outra estratégia, que pode ser combinada com a anterior, envolve a transformação e aumento de dados de treinamento para alcançar um maior nível de generalização.

No entanto, o uso destas estratégias pode causar efeitos colaterais danosos como o aumento no custo da etapa de treinamento, e mesmo assim não alcançar os resultados desejados.

Após este levantamento do cenário atual, dada uma imagem (ou conjunto de imagens) de multidão, levando em consideração toda a diversidade de configurações possíveis desta aglomeração, o número de indivíduos constituintes e suas disposições na cena, será proposto e analisado o uso de redes neurais artificiais profundas para a solução do problema da contagem por mapas de densidade que permitam estimar, com maior precisão, o número total de indivíduos na cena além de possibilitar a identificação de regiões na imagem de menor ou maior densidade.

Portanto, resolver o problema de estimativa de contagem de multidões, que é um dos desafios da área de visão computacional, utilizando técnicas de aprendizado de máquina e algoritmos de aprendizado profundo, com o menor erro esperado, é o desafio que se tentará abordar nesta dissertação de mestrado.

1.2 Contexto e Motivação

A contagem de multidões tem diferentes aplicações práticas. O monitoramento e segurança dos indivíduos diz respeito a questões de segurança pública e planejamento urbano. Pode ser considerada também questão de segurança nacional, já que países tem sofrido ataques terroristas e por isso querem monitorar seus cidadãos. Outros países que possuem governos mais rígidos querem ter maior controle sobre os indivíduos.

Os diferentes modelos e propostas encontrados na literatura têm contribuído para o crescimento e difusão da área de visão computacional. Algumas abordagens têm alcançado bons resultados e influenciado novos projetos.

Uma motivação para esse trabalho é que se possa alcançar melhores resultados com o desenvolvimento e análise de novas propostas baseadas em redes neurais artificiais profundas para a solução dos desafios que envolvem a contagem de multidões.

1.3 Objetivos

Esta Seção descreve o objetivo geral e detalha cada objetivo específico estabelecido a ser alcançado.

1.3.1 Objetivo geral

O objetivo deste trabalho é desenvolver e analisar o comportamento de redes neurais artificiais profundas, algumas delas inspiradas em novas arquiteturas de rede, na solução do problema da estimativa de contagem de multidões com mapas de densidade. Sendo assim, durante o desenvolvimento deste trabalho pretende-se responder as seguintes questões:

Questão 1. *É possível alcançar o estado da arte em contagem de multidões com uma nova arquitetura de rede em cápsulas?*

Este tipo de rede neural artificial profunda tem sido cada vez mais usado na solução dos problemas de classificação e segmentação de imagens digitais. A ideia é que seu poderoso mecanismo de roteamento dinâmico e modelagem hierárquica entre as cápsulas possa lidar melhor com a desvantagem relacionada com o número limitado de amostras de treinamento e com as limitações encontradas nas CNNs. Até o momento em que este trabalho foi desenvolvido não foram encontradas na literatura CapsNets para a solução do problema de contagem de multidões.

Questão 2. *Uma arquitetura de rede neural artifical profunda que combina as diferentes camadas da rede, utilizada originalmente em tarefas de segmentação de imagens, pode ser também utilizada na tarefa de contagem de multidões?*

Este tipo de arquitetura de rede tem alcançado excelentes resultados em tarefas de segmentação semântica de imagens digitais. Esperamos alcançar bons resultados com a combinação de camadas codificadoras e decodificadoras para capturar o contexto e permitir a localização mais precisa, que é um dos desafios deste tipo de tarefa.

Questão 3. *Um modelo híbrido, ou seja, que combina partes de uma rede CNN previamente treinada com partes de uma CapsNet é capaz de produzir mapas de densidade de alta qualidade a partir das imagens de multidões e, desta forma, solucionar o problema da contagem de multidões?*

Um modelo desta natureza pode ser mais eficiente no reconhecimento e classificação do objeto de análise, na medida que combina em sua arquitetura parte de uma CNN previamente treinada, altamente capaz de reconhecer características de nível inferior como bordas e curvas, combinando e construindo conceitos mais abstratos na parte final da rede inspirada na ideia de cápsulas.

Com base nesta suposição, cada uma das etapas seguintes estarão relacionadas ao desenvolvimento de uma arquitetura de rede e que explora a transferência de aprendizado.

Serão propostas três arquiteturas de redes neurais artificiais para a solução do problema da contagem de multidões: (i) uma arquitetura CapsNet; (ii) uma arquitetura

CNN que tenta incorporar informações de diferentes camadas da rede; e (iii) um modelo que combina parte de uma CNN previamente treinada para explorar a capacidade de transferência de aprendizado, com outra parte baseada em CapsNet.

1.3.2 Objetivos específicos

A fim de atender ao objetivo geral, foram definidos os seguintes objetivos específicos:

a) Contribuir com novas abordagens para a solução do problema de contagem de multidões baseadas em CapsNet;
b) Avalisar se a combinação de características de diferentes camadas de uma rede neural artificial profunda pode proporcionar melhores resultados;
c) Contribuir com uma abordagem híbrida que combina características de uma CNN tradicional com outra parte em CapsNets. Analisar o comportamento desta rede e qual o impacto desta combinação na solução do problema de contagem de multidões.

1.4 Justificativa

A contagem de multidões é um dos desafios atuais da visão computacional e tem diferentes aplicações práticas. Os trabalhos disponíveis na literatura utilizam diferentes abordagens e técnicas, no entanto, seus resultados ainda são passíveis de melhoria. Além disso, novas abordagens podem revelar resultados ainda não alcançados.

Existem diferentes formas de se utilizar a transferência de aprendizado na solução do problema de contagem e ainda mais formas de se desenvolver uma arquitetura de rede neural artificial profunda, formas que ainda não foram exploradas.

De acordo com a literatura recente, as CapsNets ainda não foram utilizadas na solução do problema de contagem de multidões. Neste caso, com esta abordagem, espera-se contribuir com a utilização de CapsNet na solução do problema de contagem de multidões.

Este trabalho propõe o uso de redes neurais artificias profundas para extrair e aprender características capazes de produzir mapas de densidade e caracterizar uma multidão.

1.5 Contribuições

Este trabalho tem como principal contribuição o estudo do comportamento de diferentes arquiteturas de redes neurais artificiais profundas aplicadas na solução do problema da contagem de multidões.

Estas redes analisadas foram propostas e desenvolvidas neste trabalho e combinam diferentes abordagens, sendo algumas delas novas como é o caso das CapsNet que ainda não haviam sido utilizadas neste tipo de tarefa.

Outra contribuição está relacionada a uma das arquiteturas de rede proposta neste trabalho, que tenta combinar informações de diferentes camadas da rede na tentativa de melhorar a precisão e com isso a qualidade dos mapas de densidade que serão o produto desta rede.

Uma terceira e não menos importante contribuição está no desenvolvimento e análise do comportamento de uma rede neural artificial profunda que combina parte de uma CNN previamente treinada com outra parte de uma CapsNet. Sabe-se que um dos principais motivos que deu origem as CapsNets está relacionado às limitações e problemas encontrados nas CNNs tradicionais. No entanto, ainda não está claro até que ponto podemos combinar estas ideias na construção de uma rede mais eficiente e este trabalho pode contribuir de certa forma para esclarecer essa questão.

Por fim, a comparação destas diferentes abordagens, os pontos negativos e positivos, pode contribuir para trabalhos futuros não somente, mas principalmente, na tarefa de estimativa de contagem de multidões.

1.6 Organização da dissertação

Essa dissertação está organizada da seguinte maneira. No Capítulo 2 é apresentado o referencial teórico e os trabalhos relacionados, descrevendo as principais técnicas e trabalhos utilizados na literatura recente para a contagem de multidões. No Capítulo 3 é descrita a metodologia utilizada para o desenvolvimento da pesquisa e detalhadas cada uma das propostas e suas respetivas arquiteturas. No Capítulo 4 são apresentados os experimentos e os resultados alcançados com cada uma das propostas. Por fim, no Capítulo 5 são apresentadas as conclusões e os trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este Capítulo apresenta os principais conceitos teóricos e trabalhos relacionados com a proposta desta dissertação.

2.1 Imagem Digital

Pedrini e Schwartz (2007) definiram uma imagem como uma função $f(x, y)$, sendo o valor fornecido pelas coordenadas espaciais (x, y) a intensidade ou o brilho da imagem naquele determinado ponto. Neste contexto, a intensidade luminosa é o produto entre a quantidade de luz incidente na cena (iluminância) e a quantidade de luz refletida pelos objetos em cena (reflectância). Em resumo, a Equação 2.1 apresenta a definição da função $f(x, y)$, de modo que i representa a iluminância e r a reflectância.

$$f(x, y) = i(x, y)r(x, y) \tag{2.1}$$

Cada posição ou elemento desta matriz é conhecido como *pixel*, do inglês *picture element*, e tem um papel importante pois representa o brilho (ou nível de cinza) naquele ponto. É a menor unidade em uma imagem.

Uma imagem pode também ser analisada e processada levando em consideração outras propriedades como a relação de adjacência e vizinhança entre os pixels. Este relacionamento pode ser explorado de acordo com cada problema abordado.

Uma imagem digital, segundo Pedrini e Schwartz (2007), pode ser obtida a partir do processo de digitalização pelas etapas de amostragem e quantização. A etapa de amostragem é responsável por discretizar as coordenadas (x, y), do domínio da definição de imagem, para uma matriz de $(M \times N)$ elementos, onde M corresponde ao eixo x e N ao eixo y da mesma imagem.

Em imagens monocromáticas, a etapa de quantização se refere a definição do valor inteiro L dos níveis de cinza para cada um dos *pixels* da imagem que podem, comumente, assumir valores no intervalo $[0, 255]$.

Para as imagens coloridas, a quantização pode ser realizada atribuindo a cada *pixel* os níveis de vermelho, verde e azul, formando três componentes RGB (*Red, Green and Blue*). Na Figura 6, tem-se uma imagem RGB e outra em escala de cinza.

Figura 6 – Imagem no modelo RGB (a); e em escala de cinza (b).

(a) **(b)**

Fonte: Elaborada pelo autor

O tamanho de uma imagem digital pode variar de acordo com a aplicação. Uma imagem RGB com dimensões espaciais de 32 × 32 pixels e 3 canais (RGB), tem um número total de pixels na imagem igual a 32 × 32 × 3 (3072 pixels).

Existem ainda outras propriedades como a conectividade, que é muito importante para estabelecer os limites dos objetos e componentes de regiões semanticamente relevantes em uma imagem.

Precisamos de ferramentas capazes de extrair significado a partir destes dados. Neste caso, utilizamos o aprendizado de máquina na tentativa de fazer com que uma máquina possa aprender a reconhecer padrões e extrair significados a partir de imagens digitais.

2.2 Aprendizado de máquina e reconhecimento de padrões

Os seres humanos e boa parte dos animais tem natural facilidade e habilidade em algumas tarefas que são extremamente difíceis para uma máquina. Em uma tarefa de reconhecimento ou classificação de objetos, por exemplo, o cérebro humano processa informações visuais principalmente no espaço semântico, ou seja, extraindo características semanticamente significativas, como segmentos de linha, limites, formas e assim por diante.

Os sistemas computacionais encontram dificuldade em processar informações visuais com a mesma capacidade dos animais e seres humanos. Eles precisam processar informações visuais no espaço de dados, formado por recursos que possam ser detectáveis, mas menos significativos como um todo, como cores, texturas, etc.

Portanto, a metodologia de processamento é bem diferente entre computadores e seres humanos. No entanto, existe algo em comum: da mesma forma que os humanos precisam ser treinados para desempenhar uma determinada tarefa, os computadores também podem ser treinados, obviamente de maneiras diferentes.

Este treinamento é realizado por meio de amostras. Por exemplo, em tarefas da área de visão computacional, um modelo de aprendizado de máquina pode ser treinado com várias imagens de exemplo de um objeto e delas extrair características como forma e cor que permitiram o reconhecimento posterior.

O reconhecimento de padrões, de acordo com Haykin (2009), é definido como o processo pelo qual um padrão/sinal recebido é atribuído a uma classe dentre um número pré-determinado de classes, rótulos ou categorias.

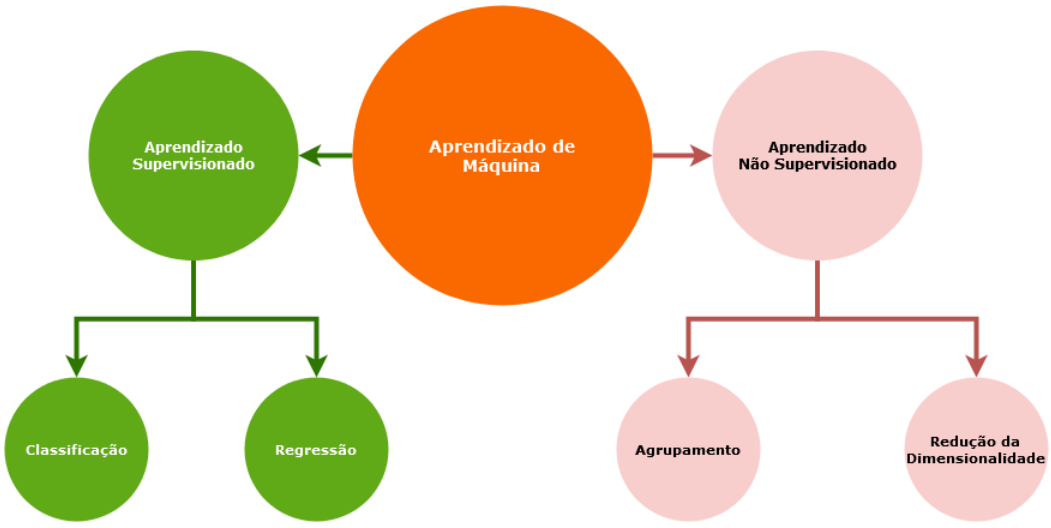
Segundo Kuncheva (2014), os métodos de aprendizado de máquina podem ser divididos em duas grandes categorias de aprendizado: (i) não supervisionado; e (ii) supervisionado. Na Figura 7 algumas variações destas categorias.

No aprendizado supervisionado as amostras são rotuladas. Este rótulo ou variável associada deverá ser utilizada na tarefa de regressão (quando a variável é um valor contínuo) ou de classificação (quando a variável de saída é um rótulo de classe). Neste tipo de aprendizado o objetivo é determinar um mapeamento de x para y, dado um conjunto de treinamento contendo pares (x_i, y_i), em que $y_i \in Y$ é denominado de rótulo ou objetivo do exemplo x_i.

Já no aprendizado não supervisionado, as amostras não são rotuladas. Portanto, o objetivo é tentar encontrar padrões de comportamento no conjunto de dados não rotulados. Dessa forma, os algoritmos utilizados nessa abordagem devem encontrar entre os dados padrões representativos e que possibilitem a divisão da base em grupos.

Em Russell e Norvig (2010), foi definida mais uma nova categoria denominada

Figura 7 – Categorias e ramificações mais comuns para o aprendizado de máquina.



Fonte: Elaborada pelo autor

de aprendizado semi-supervionado. Esta categoria é uma combinação do aprendizado supervisionado e não supervisionado que se aplica em cenários com uma quantidade baixa de amostras rotuladas e uma quantidade relativamente maior ou equivalente de amostras não rotuladas. Em algum momento, durante a fase de aprendizado, os rótulos serão utilizados.

O processo de aprendizado como um todo será responsável pela formação da memória. Os padrões armazenados em memória (seja qual for a representação desta memória) devem ser recordados e explorados de acordo com as experiências passadas, permitindo o uso na solução de novos problemas dentro do mesmo domínio na qual foi treinado. Em alguns casos, é possível extrapolar e reutilizar esta memória ou modelo na solução de novos problemas (*Transfer Learning*).

Os procedimentos para aprendizado de máquina são expressos, geralmente, em forma de algoritmo. Em tarefas mais desafiadoras, estes algoritmos devem proporcionar um aprendizado mais profundo. As redes neurais artificiais surgem como modelos computacionais complexos, porém eficientes, capazes de permitir aos computadores o aprendizado de máquina mais profundo.

2.3 Redes Neurais Artificiais

Uma Rede Neural é um algoritmo de otimização bioinspirado que funciona a partir de um conjunto de unidades de processamento ou células de computação organizadas em camadas e que possui aplicações práticas. O tipo e modelagem do problema pode influenciar na arquitetura da rede, na quantidade e na maneira como as unidades de processamento são conectadas.

Para Haykin (2009), uma rede neural é um processador maciçamente e paralelamente distribuído, construído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso, e que se assemelha ao cérebro em dois aspectos:

1) O conhecimento é adquirido pela rede a partir de seu ambiente com um processo de aprendizado;
2) Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

Uma rede neural deverá armazenar o conhecimento sobre um fenômeno físico ou ambiente de interesse codificado através do treinamento.

De acordo com Pedrini e Schwartz (2007), uma rede neural artificial é um modelo de grafo orientado em que os nós representam neurônios artificiais e as arestas orientadas

denotam as conexões entre as entradas e as saídas dos neurônios.

O *perceptron* é um classificador linear criado por Rosenblatt (1958), representado como um modelo de rede neural artificial de arquitetura simples que foi desenvolvido para lidar com o problema de reconhecimento de padrões.

O classificador de um neurônio apenas ilustrado na Figura 8, separa duas classes utilizando uma função discriminante linear, de modo que os vetores de uma classe obtêm uma saída de valor 1 e os da outra classe o valor 0. O modelo é iniciado com pesos aleatórios w que serão corrigidos iterativamente no processo de treinamento supervisionado.

As redes *perceptron* de multicamadas, do inglês *Multilayer Perceptron* (MLP), introduzidas por Rumelhart, Hinton e Williams (1986), são uma melhoria em relação aos *perceptron* originais, fornecendo uma capacidade de classificar dados não separáveis linearmente.

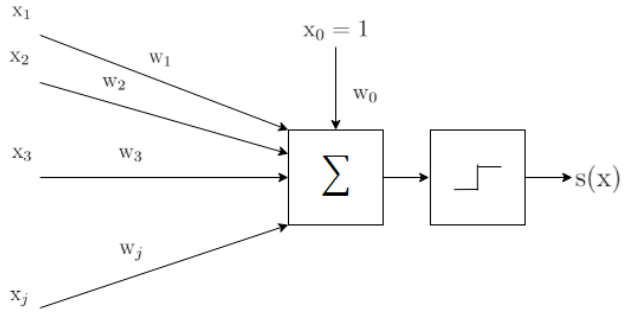
MLPs são modelos de aprendizado com objetivo de aproximar alguma função f. Por exemplo, um classificador $y = f(x)$ mapeia uma entrada x para uma categoria y. Este tipo de rede define um mapeamento $f(x;\theta)$ e aprende o valor dos parâmetros θ que resultam na melhor aproximação da função.

Em arquiteturas multicamadas, a saída de uma camada é utilizada como entrada para a camada seguinte e uma função de ativação não-linear é aplicada, permitindo assim que funções complexas sejam obtidas para o mapeamento entre a entrada e a saída. Na Figura 9 temos um exemplo de uma arquitetura básica de uma rede multicamadas.

Não há ligações entre nós da mesma camada e camadas não adjacentes não são conectadas diretamente. Todos os nós das camadas escondidas geralmente possuem a mesma função de ativação.

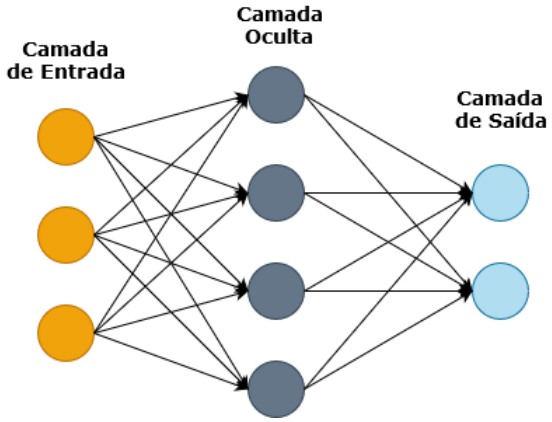
O treinamento destas redes é realizado com o algoritmo *backpropagation*, que aplica o método do gradiente descendente para minimizar a função de erro que depende da

Figura 8 – Modelo de rede neural composto de uma unidade perceptron.



Fonte: Elaborada pelo autor

Figura 9 – Modelo básico de rede neural artificial multicamadas com camada de entrada, camada oculta e camada de saída.



Fonte: Elaborada pelo autor

comparação do resultado obtido na saída com o valor esperado a partir das amostras utilizadas na etapa de treinamento.

As redes *feedforward* são de extrema importância no aprendizado de máquina e são base das muitas aplicações comerciais importantes. Como exemplo, as redes neurais convolucionais que são muito utilizadas em tarefas de reconhecimento e classificação de objetos a partir de imagens digitais.

2.4 Redes Neurais Artificiais Convolucionais

A motivação inicial para as CNNs tem origem no trabalho de Hubel e Wiesel (1959), que exploraram o córtex visual de um gato e encontraram pequenas regiões com células sensíveis a regiões específicas no campo visual, sugerindo a existência de um campo receptivo que é um conceito usado para descrever a ligação entre partes dos campos visuais e neurônios individuais.

A existência de neurônios corticais especializados em certas orientações revela que a excitação de determinadas células depende da forma e orientação dos objetos e das suas características no campo visual. Por exemplo, linhas verticais fazem com que algumas células neuronais sejam excitadas, enquanto linhas horizontais excitam outras. Esta descoberta sugere que os mamíferos usam diferentes camadas e a combinação delas para construir partes de imagens em diferentes níveis de abstração, princípio semelhante ao da extração hierárquica de características, que se baseia na ideia de que as células são conectadas usando uma arquitetura em camadas.

Com base nessa inspiração biológica, um dos primeiros modelos de rede neural artificial a simular este comportamento foi o *Neocognitron* (FUKUSHIMA; MIYAKE,

1982). Porém, existem diferenças entre esse modelo e as modernas CNNs, sendo a mais importante delas a noção de compartilhamento de pesos.

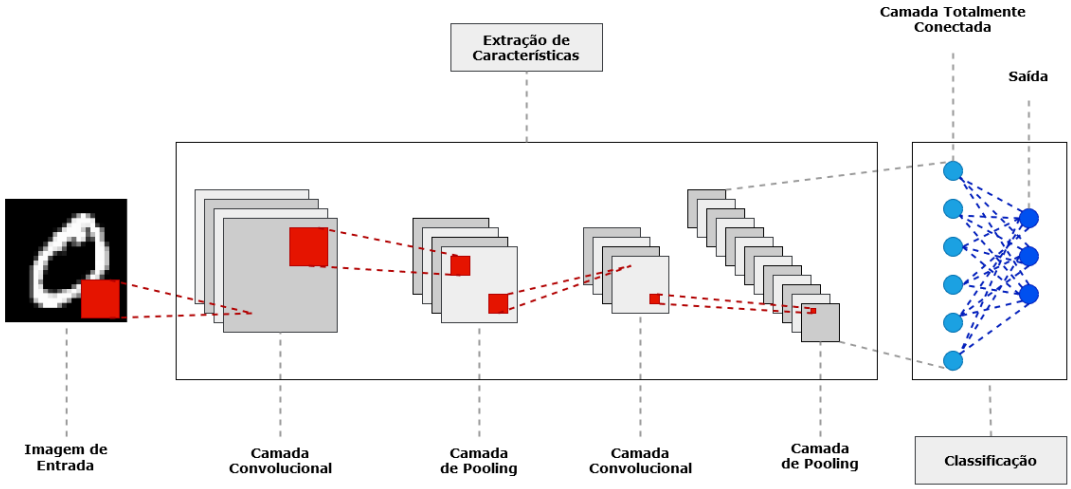
Uma das primeiras arquiteturas totalmente convolucionais foi a rede *LeNet-5* (LECUN et al., 1998). Essa rede teve aplicações práticas, pois foi utilizada para identificar dígitos manuscritos de códigos postais, além de servir de inspiração para vários novos trabalhos.

As CNNs foram uma das primeiras histórias de sucesso do aprendizado profundo. Ganhou mais notoriedade com o sucesso alcançado na competição ImageNet (RUSSAKOVSKY et al., 2015). Em 2012, uma CNN chamada AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) foi a vencedora da competição. Ela obteve alto desempenho na época e mesmo que fosse um modelo caro computacionalmente, foi viabilizado devido a utilização de unidades de processamento gráfico, do inglês *Graphic Processing Unit* (GPU), durante o treinamento.

CNNs são amplamente utilizadas em tarefas de visão computacional, na qual o processamento é realizado por uma sequência de camadas que combinadas são utilizadas para modelar padrões visuais em diferentes níveis de abstração, sendo as primeiras responsáveis por capturar *features* de baixo nível como bordas e contornos e as seguintes de mais alto nível na definição das formas. Um exemplo de arquitetura básica de uma CNN está ilustrada na Figura 10.

Um dos segredos para o sucesso de uma arquitetura de rede neural artificial reside em adaptar e projetar cuidadosamente a sua estrutura para lidar com a compreensão semântica do domínio em questão.

Figura 10 – Arquitetura básica de uma CNN usada para classificação.



Fonte: Elaborada pelo autor

Algumas camadas são mais comuns nas CNNs: camada de convolução ou convolutivas, camadas de ativação e camadas de agrupamento (*subsampling*, *downsampling* ou *pooling*). Além destas, um conjunto final de camadas geralmente era totalmente conectada *(fully-connected)* e mapeada de maneira específica para um conjunto de nós de saída que mais se correlacionam com a classe ou rótulo do objeto.

A convolução é uma operação linear realizada entre duas funções e que tem diversas aplicações em processamento de sinais. É a principal operação e o coração deste tipo de rede que inclusive recebe o nome por conta dela. Uma convolução entre duas funções *f(x)* e *g(x)* é denotada por *f(x)*g(x)* e definida pela integral dada pela Equação 2.2:

$$f(x) * g(x) = \int_{-\infty}^{\infty} f(\alpha) g(x - \alpha) d\alpha \tag{2.2}$$

em que α é uma variável de integração. Uma operação de convolução discreta 2D (2 Dimensões) é aplicada sobre duas funções bidimensionais *f(x, y)* e *g(x, y)*, representada como matrizes discretas de dimensão A × B e C × D respectivamente, com periodicidade em um determinado período M e N nas direções x e y, de acordo com a Equação 2.3 e Equação 2.4.

$$M \geq A + C - 1 \tag{2.3}$$

$$N \geq B + D - 1 \tag{2.4}$$

A operação de convolução discreta bidimensional é definida pela Equação 2.5:

$$f(x, y) * g(x, y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m, n) g(x - m, y - n) \tag{2.5}$$

para x = 0, 1, 2, ..., M - 1 e y = 0, 1, 2, ..., N -1. A matriz M × N da equação 2.5 é um período da convolução discreta bidimensional que estará livre de interferência de outros períodos adjacentes se os períodos forem escolhidos de acordo com as Equações 2.3 e 2.4.

Os estados em cada camada são organizados de acordo com uma estrutura de grade espacial e os relacionamentos espaciais são herdados de uma camada para a próxima, pois cada valor é baseado em uma pequena região espacial local na camada anterior, definida por uma janela ou *kernel*.

A operação de convolução e a transformação para a próxima camada são estritamente dependentes dessas relações e tem como produto o mapa de ativação ou mapa de atributos.

As funções de ativação introduzem um componente não linear nas redes neurais, que permitem aprender mais do que relações lineares ao longo do sinal de entrada. É o mecanismo matemático implementado nas unidades de processamento para tomar a decisão de passar ou não o sinal adiante. As funções mais comuns são: *sigmóide*, tangente hiperbólica *(tanh)* e ReLU.

A função ReLU é uma das mais utilizadas pois trata-se de um mapeamento simples e individual dos valores de ativação que tem enormes vantagens sobre outras funções de ativação saturantes como *sigmóide* e *tanh*, tanto em termos de velocidade quanto de precisão (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

As camadas de convolução são geralmente intercaladas ou combinadas por camadas de *pooling* e por funções de ativação, não necessariamente nesta ordem.

As camadas de *pooling* realizam a redução da amostragem de mapas de recursos compactando os dados para extrair somente informações supostamente mais relevantes, ajudam a tornar a representação aproximadamente invariável para pequenas traduções da entrada e além de controlar o número de parâmetros.

Para Riesenhuber e Poggio (1999), o modelo original de uma célula complexa pode ser vista como um agrupamento de entrada de uma matriz de células simples em diferentes locais para gerar sua resposta invariável na posição. Um componente chave dos modelos hierárquicos de processamento cortical, como os de Fukushima e de Hubel e Wiesel, está no *pooling* que visa aumentar a invariância da resposta neuronal de certas transformações de estímulos (RIESENHUBER; POGGIO, 2002).

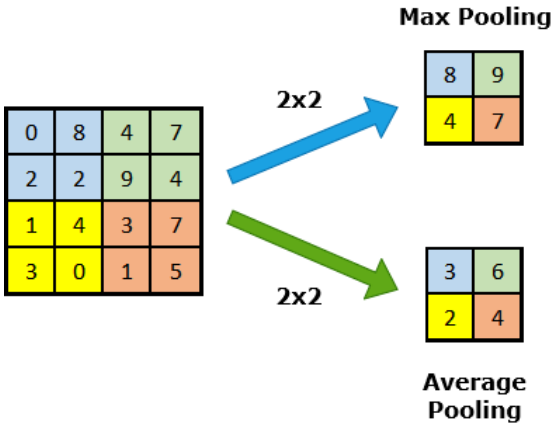
A ideia básica do modelo hierárquico esboçado por Perrett e Oram (1993) era a de que a invariância a qualquer transformação (não apenas as transformações no plano da imagem, como no caso do Neocognitron) poderia ser construída através de *pooling* e em várias versões transformadas do mesmo estímulo.

Os tipos de operações mais comuns nesta camada são o *Max-Pooling* que extrai o maior elemento da região e *Average-Pooling* que extrai o valor médio da região.

As camadas de *pooling* oferecem uma quantidade de invariância translacional em cada nível, no entanto a localização precisa das características mais ativas é descartada. Além disso, há uma redução significativa do número de entradas que são enviadas para as próximas camadas. Em domínios maiores, seria mais interessante ter mais *features* nas próximas camadas.

A Figura 11 apresenta um exemplo de agrupamento máximo e médio e, em ambos os casos, a resposta é invariante da posição. É permitido também realizar a operação de soma, conhecida como *Sum-Pooling*, nestas camadas.

Figura 11 – Exemplo da operação de agrupamento máximo *(Max-Pooling)* e médio *(Average-Pooling)*.



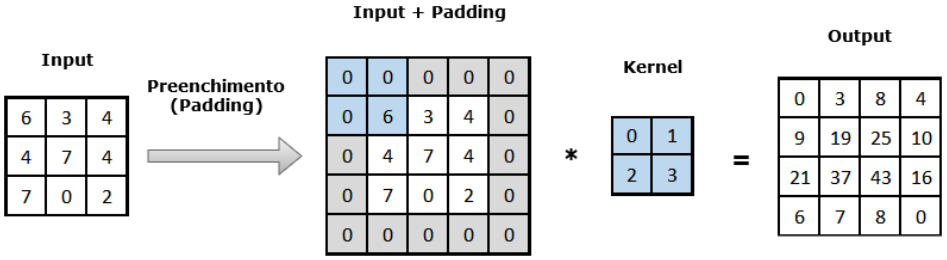
Fonte: Elaborada pelo autor

Um outra maneira de reduzir a área espacial da imagem, como alternativa ao *pooling*, desta vez pela operação de convolução, está na configuração da "passada", do inglês *stride*, que reduz o nível de granularidade da convolução.

As CNNs são configuradas e inicializadas com valores ou estruturas comuns chamadas de hiperparâmetros, que serão utilizados no processamento da rede e seu ajuste contribui significativamente para os resultados esperados, sendo os mais comuns: taxa de aprendizado, função de perda, número de camadas, inicialização dos pesos, fator de preenchimento *(padding)*, o viés e a regularização.

A redução que ocorre na camada de convolução em geral não é desejável pois tende a perder algumas informações ao longo das bordas da imagem (ou do mapa de características, no caso de camadas ocultas). Esse problema pode ser resolvido usando um fator de preenchimento, do inglês *padding*. Na Figura 12 temos um exemplo do fator de preenchimento.

Figura 12 – Exemplo do uso de fator de preenchimento *(padding)*, representado por 0s (zeros) na borda da imagem (cor cinza).



Fonte: Elaborada pelo autor

Uma CNN geralmente é inicializada com pesos aleatórios que serão alterados ao longo da etapa de treinamento. Quando é utilizada a transferência de aprendizado, os pesos de uma rede que já foi previamente treinada são reutilizados na inicialização de uma nova rede e serão ajustados durante a etapa de treinamento.

Outro tipo de hiperparâmetro que merece destaque diz respeito ao uso de uma camada de regularização, que elimina ou desativa um subconjunto de neurônios da rede, reduzindo ajustes excessivos na tentativa de melhorar a sua capacidade de generalização. A camada de regularização mais comum e que foi utilizada em uma das propostas deste trabalho foi a *Dropout*.

Desde sua origem, os modelos baseados em CNN têm se diferenciado principalmente em relação ao número de camadas, de funções de ativação mais estáveis, como o ReLU, e em relação ao emprego de estratégias na etapa treinamento visando melhorar os resultados.

Porém, algumas críticas são atribuídas às CNNs como, por exemplo, o fato de que elas possuem poucos níveis de estrutura. Além disso, não são treinadas para reconhecer o relacionamento existente entre as partes que constituem o objeto de análise.

Outra crítica, que motivou novas abordagens (SABOUR; FROSST; HINTON, 2017), é direcionada ao *pooling*, que desconsidera informação da localização das *features* extraídas, pela maneira como se dá a ativação e resposta a determinados filtros e o uso excessivo contribui para a perda de informações pelo simplificação excessiva, o que dificulta a localização de objetos menores em determinadas tarefas. CNNs são, em geral, dependentes das operações que ocorrem nestas camadas.

Uma das estratégias para lidar com estas limitações é o aumento das amostras de treinamento, pois dois objetos idênticos em diferentes orientações não são representados da mesma maneira. Outra estratégia envolve a ampliação da arquitetura da rede, empilhando mais camadas. Ambas as estratégias podem causar como efeito colateral uma maior complexidade e um alto custo principalmente na etapa de treinamento.

É possível também relacionar o sucesso de algumas abordagens baseadas em CNN no fato de que elas possam estar memorizando as imagens de treinamento ao invés de estarem extraindo características mais genéricas e relevantes.

Trabalhos recentes mostraram que algumas redes neurais artificiais podem ter seu resultado prejudicado com a alteração de apenas um único pixel na entrada (SU; VARGAS; SAKURAI, 2019). No entanto, o ataque é mais difícil nas CNNs por alcançarem um pouco mais de precisão e confiança na classificação.

De qualquer maneira, como as CNNs tentam tornar as atividades neurais invariáveis a pequenas mudanças de ponto de vista, não se sabe ainda qual é o menor número

de alterações de pixel ou do objeto de análise são necessários para “enganar” uma rede neural ou comprometer o seu resultado.

Novas propostas têm surgido como alternativa as CNNs tradicionais, na tentativa de lidar de forma mais eficiente com a relação espacial e hierárquica entre as partes do objeto de análise e evitar a simplificação excessiva, sendo uma delas as CapsNets.

2.5 Redes Neurais Artificiais em Cápsulas

As CapsNets ou Redes Neurais em Cápsulas simbolizam uma novidade recente nas arquiteturas de redes neurais artificiais profundas, um modelo biologicamente plausível que alcançou excelentes resultados no conjunto de dados MNIST (LECUN et al., 1998), um feito tradicionalmente realizado por CNN.

Para Sabour, Frosst e Hinton (2017), cápsulas são grupos de neurônios semanticamente significativos e representam o principal bloco da arquitetura CapsNet. Sua proposta é de um funcionamento mais semelhante ao do cérebro humano, principalmente do subsistema de processamento visual, levando em consideração não somente a presença das características na ativação, mas também uma determinada organização destas características.

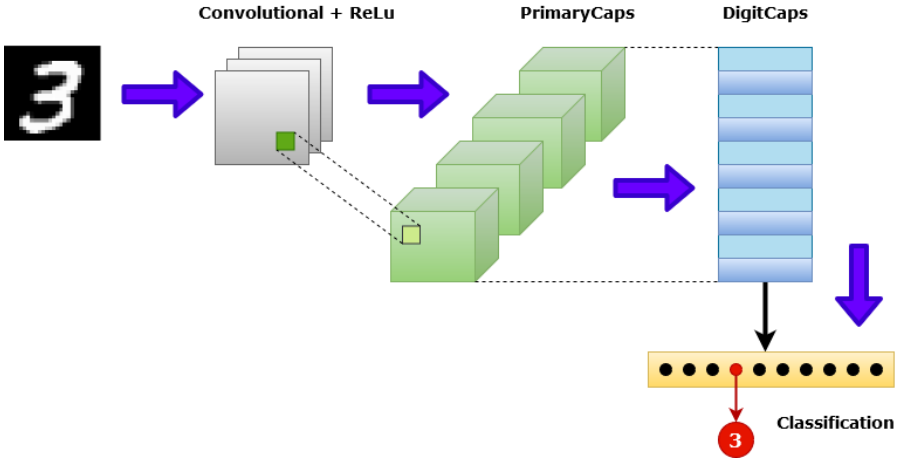
A proposta é de fornecer resultados mais compactos, cujos elementos ou partes constituintes são ou estão firmemente unidos entre si, e cada cápsula representa a presença (ou a probabilidade da presença) e também os parâmetros de instanciação de partes de uma entidade que a cápsula será treinada para detectar.

A arquitetura básica de uma CapsNet está ilustrada na Figura 13 e consiste basicamente das seguintes camadas:

a) Camada Convolucional: extrair algumas características básicas da imagem de entrada, como bordas ou curvas;
b) Camada de Cápsulas Primárias *(PrimaryCaps)*: tenta extrair e combinar *features* um pouco mais complexas a partir das convoluções, respeitando as relações todo-parte;
c) Camada de Dígitos *(DigitCaps)*: camada de mais alto nível que contém todos os parâmetros de instanciação para a previsão final, onde o comprimento do vetor é a confiança do objeto encontrado.

As CapsNets apontam para uma forma de roteamento mais efetiva do que a “forma primitiva” e supostamente ruim realizado nas camadas de *pooling*, de forma que cada parte da entrada seja direcionada a neurônios ou cápsulas que saibam lidar com ela, tentando respeitar uma hierarquia entre as partes semanticamente relacionadas.

Figura 13 – Exemplo de uma arquitetura básica CapsNet.



Fonte: Elaborada pelo autor

Esta proposta responde à questão de "como entidades visuais maiores e mais complexas podem ser reconhecidas usando combinações das poses previstas por cápsulas ativas de nível inferior".

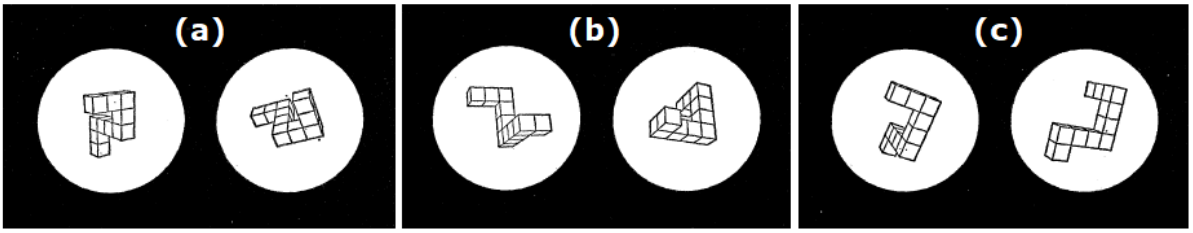
O trabalho das cápsulas é fazer a renderização inversa da imagem, o que significa que obtemos os parâmetros de instanciação, como ângulo, escala e posição do objeto, analisando o objeto de acordo com as amostras de treinamento fornecidas. Esse é um tipo de computação muito diferente do que juntar partes instanciadas para criar conjuntos familiares, e é justamente esta característica que tornam boas as cápsulas (SABOUR; FROSST; HINTON, 2017).

Considerando que mudanças no ponto de vista levam a mudanças correspondentes nas atividades neurais, as cápsulas consideram uma melhor estratégia lidar com a equivariância e não com a invariância na tradução. Em uma CNN, por exemplo, o rótulo final é invariante do ponto de vista, ou seja, um objeto é identificado mas perde-se a informação do ângulo de rotação, impedindo que ele seja corretamente classificado caso sofra alterações.

A hipótese de Shepard e Metzler (1971), no trabalho que estudaram a rotação mental em que candidatos foram levados a analisar desenhos e classificá-los, como ilustra a Figura 14, era de que a tarefa seria realizada formando uma imagem mental tridimensional de um dos objetos representados e girando toda a imagem, na imaginação, para ver se ela poderia ser correspondida com a outra imagem.

Este trabalho sugere a existência de um relacionamento de partes com o todo na definição e ou identificação dos objetos de análise no processamento visual, como se fosse necessário, em alguns casos, girar ou modificar a imagem mental da representação do objeto, propriedade descartada ou pouco explorada em redes do tipo CNN.

Figura 14 – Exemplos de pares de desenhos de formas com variação de perspectiva (SHEPARD; METZLER, 1971). Um par idêntico que difere em uma rotação de 80º no plano da imagem (a); um par idêntico que difere em uma rotação de 80º em profundidade (b); e um par diferente que não pode ser trazido à congruência por qualquer rotação (c).



Fonte: Adaptada de (SHEPARD; METZLER, 1971)

Se considerarmos a utilização de uma CNN para o experimento de Shepard e Metzler (1971), mesmo se ela for treinada com imagens em diversas rotações possíveis do objeto de análise, ela encontrará dificuldade em generalizar a representação, podendo comprometer os resultados.

Em um cenário onde os dados para treinamento são limitados e contém apenas imagens do ângulo frontal dos objetos, as CapsNets podem ajudar a generalizar melhor para novos pontos de vista de forma mais eficiente. No entanto, podem levar mais tempo de treinamento para realizar esta tarefa por conta do seu algoritmo de roteamento dinâmico que é mais lento na definição do acordo entre as cápsulas de diferentes níveis.

Uma cápsula detecta um tipo específico de objeto ou parte de objeto e produz, principalmente:

1) a probabilidade de um objeto do tipo desejado estar presente;
2) a pose generalizada do objeto que inclui posição, orientação, escala, etc.

A saída de uma cápsula é calculada usando uma função de compressão não linear chamada de *squashing*, que comprime os valores para um intervalo entre 0 e 1, mantendo as proporções de seus parâmetros, permitindo que seja representada a probabilidade da presença da entidade.

O cálculo é realizado pela Equação 2.6 em que v_j é o vetor de saída da j-ésima cápsula na camada atual e s_j é seu vetor de entrada.

$$v_j = \frac{||s_j||^2}{1 + ||s_j||^2} \times \frac{s_j}{||s_j||} \tag{2.6}$$

Para todas, exceto a primeira camada de cápsulas, a entrada total de uma cápsula s_j é uma soma ponderada de todos os vetores de previsão q_i das cápsulas na camada

anterior e é produzida pela multiplicação de $q'_{j|i}$ da saída de uma cápsula na camada abaixo por uma matriz de peso W_{ij} que são obtidos, respectivamente, pela Equação 2.7 e pela Equação 2.8:

$$q'_{j|i} = W_{ij} q_i \tag{2.7}$$

$$s'_j = \sum c_{ij} q'_{j|i} \tag{2.8}$$

na qual $q'_{j|i}$ é um vetor de previsão produzido pela transformação da saída q_i de uma cápsula na camada abaixo em um peso W_{ji}, e c_{ij} é o coeficiente de acoplamento determinado pelo processamento de roteamento dinâmico iterativo.

Parte da compressão atribui um comprimento unitário ao vetor e a função *softmax* garante que a soma dos coeficientes do acoplamento c_{ij} permaneça na faixa correta, cuja entrada b_{ij} é a probabilidade logarítmica anterior de acoplamento entre as cápsulas, dada pela Equação 2.9.

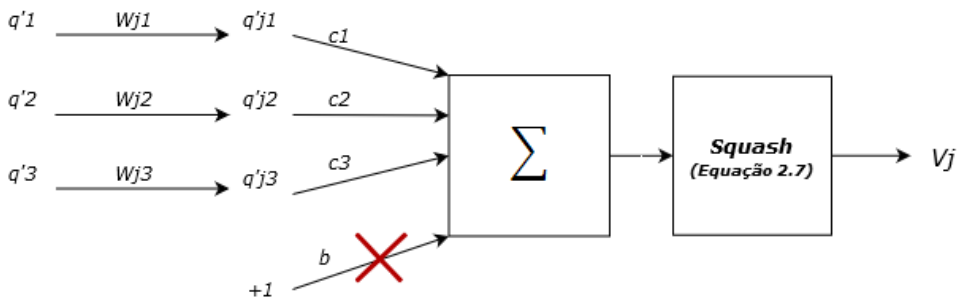
$$c_{ij} = \frac{exp(b_{ij})}{\sum_k exp(b_{ik})} \tag{2.9}$$

O algoritmo de roteamento dinâmico *(Dynamic Routing Algorithm)*, detalhado no Algoritmo 1, viabiliza a comunicação entre as cápsulas, levando em consideração o acordo entre elas, criando representações que permitem à rede obter equivariância e relações significativas de parte para o todo, pouco comuns nas CNNs. Na

Figura 15 temos um exemplo de um diagrama de uma cápsula. Neste modelo não há viés explícito como entrada, mas ele pode estar incluido nas matrizes W_{ij}.

O algoritmo de roteamento recompensa o acordo entre a saída real da cápsula v_j e o vetor previsto $q'_{j|i}$, na qual cada cápsula de uma camada anterior seleciona uma outra cápsula na próxima camada como pai e envia sua ativação para ela.

Figura 15 – Diagrama de uma cápsula modelada como um neurônio artificial.



Fonte: Elaborada pelo autor

Cápsulas ativas em um nível fazem previsões por suas matrizes de transformação para os parâmetros de instanciação de cápsulas de nível superior. Quando várias previsões concordam, uma cápsula de nível superior se torna ativa.

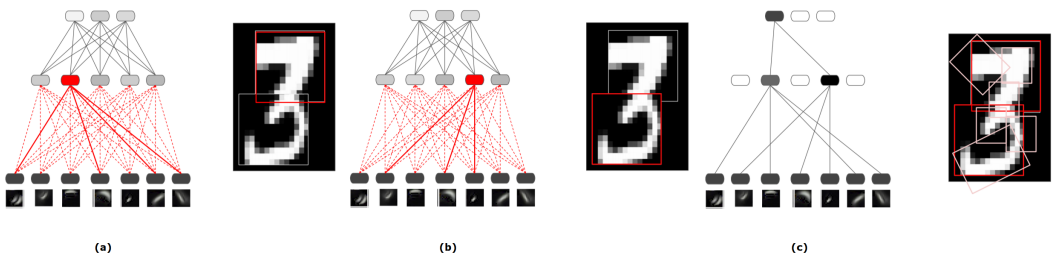
Algorithm 1 Algoritmo de Roteamento Dinâmico por Concordância

1. **procedimento** Roteamento($q'_{j|i}, r, l$)
2. para todas as capsulas i da camada l e cápsulas j na camada $(l+1)$: $b_{ij} \leftarrow 0$.
3. **Para** r iterações **Faça**
4. Para toda cápsula i na camada l : $c_i \leftarrow softmax(\mathrm{b}_i)$ (Equação 2.9)
5. Para toda cápsula j na camada $(l+1)$: $s_j \leftarrow \sum c_{ij} q'_{j|i}$
6. Para toda cápsula j na camada $(l+1)$: $v_j \leftarrow squashing(\mathrm{s}_j)$ (Equação 2.6)
7. Para toda cápsula i na camada l e para toda cápsula j na camada $(l+1)$: $b_{ij} \leftarrow \mathrm{b}_{ij} + (q'_{j|i}.v_j)$
8. **Retorne** v_j

A cápsula de nível superior tenta encontrar um subconjunto das previsões que concordam, obtendo uma pontuação alta se muitas previsões concordarem. Um objeto existe se houver concordância entre previsões de várias partes, conforme ilustra o exemplo da Figura 16.

A classificação correta ou saída esperada é representada por uma quantidade de neurônios ativos que capturam diferentes aspectos do relacionamento entre as partes do objeto e não somente por um único neurônio ou um conjunto de neurônios de código "grosseiro" que desconsideram estas propriedades.

Figura 16 – Exemplo dos acordos entre cápsulas para a classificação de um dígito. As linhas vermelhas indicam um "acordo" na identificação da parte superior do dígito (a); outra cápsula que "concorda" na identificação da parte inferior do dígito (b); e a hierarquia das cápsulas na classificação final (c).



Fonte: Elaborada pelo autor

O roteamento dinâmico é realizado entre duas camadas adjacentes de cápsulas, de nível superior e inferior, e o roteamento entre um par de camadas deve ser concluído antes de iniciar o roteamento entre o próximo par de camadas.

Se uma parte do objeto detectada por camadas de nível inferior se mover para uma posição muito diferente, ela será representada por uma cápsula. Porém, se for apenas uma pequena distância, ela será representada pela mesma cápsula, mas as saídas da pose da cápsula serão alteradas.

O processo de roteamento tem uma forte semelhança com o ajuste de uma mistura de gaussianos usando maximização de expectativas, do inglês *Expectation-Maximization* (EM), onde as cápsulas de nível superior desempenham o papel dos gaussianos e os meios das cápsulas ativadas de nível inferior para uma única imagem de entrada desempenham o papel dos pontos de dados (HINTON; SABOUR; FROSST, 2018).

Com as CapsNets ainda é possível explorar a capacidade de reconstrução do objeto de análise. A ideia é definir um decodificador simples que produza uma imagem adicionando contribuições de cada cápsula, na qual um conjunto de cápsulas semanticamente relevante aprende um "modelo" fixo que permitirá sua reconstrução.

A reconstrução pode ser utilizada como validação, de forma que o modelo é classificado com base em quão perto a reconstrução corresponde à imagem original.

Para lidar com o problema de variação de escala, perda de informações espaciais e problemas com conjuntos de treinamento limitados, também como alternativa as CNNs tradicionais, propomos o uso de diferentes arquiteturas em CapsNets que permitam estimar, com maior precisão, o número de indivíduos na multidão.

2.6 Contagem de Multidões

O problema da contagem de multidões, apesar de ser atualmente tratado como um desafio da área de visão computacional, tem suas origens na década de 60 com o trabalho de Jacobs (1967). Insatisfeito com os métodos de contagem da época, ele desenvolveu um método de contagem de multidões baseado em critérios objetivos, que se resume na medida total da área ocupada multiplicada pela estimativa de densidade da ocupação. Seidler, Meyer e Gillivray (1976) propuseram o conceito de estimativa por zoneamento, baseado na contagem direta de amostras em diferentes pontos.

Zeitz et al. (2009) destacam em seu trabalho a experiência no envolvimento com a multidão, os aspectos psicológicos e sociais, de planejamento e de organização dos eventos que envolvem multidões, dentre outros fatores. Pudemos obter intuições a partir deste trabalho capazes de contribuir sensivelmente com nossas pesquisas.

Com o avanço da computação, surgiram algoritmos computacionais capazes de realizar a estimativa de contagem por meio de imagens digitais, baseando-se em amostras locais e globais da multidão.

Levando em consideração a relevância do tema, algumas pesquisas discutem sobre a análise de multidões e foram motivadas, principalmente, pelo avanço da tecnologia e pelo estudo do comportamento humano, como em (LI et al., 2008) e (JUNIOR; MUSSE; JUNG, 2010).

Estas pesquisas discorrem sobre a necessidade do monitoramento e vigilância de multidões, através da análise dos indivíduos que a constituem (caso sejam passíveis de identificação) ou análise da multidão como se fosse uma entidade única.

Para Li et al. (2008), é importante pensar em uma multidão pela perspectiva de sua densidade, onde ela pode se apresentar, em geral, de maneira pouco, média ou muito densa. Esta característica é importante pois o modelo produzido deverá contemplar este comportamento e ter capacidade de generalização, independente desta variação.

No entanto, este aspecto é de fundamental importância pois ele também contribuirá na definição da técnica escolhida. Por exemplo, não podemos utilizar uma técnica baseada em descritores de pessoas (cabeça, ombros e corpo) em cenas de alta densidade, onde a oclusão e variação do ponto de vista não permitem a identificação isolada dos indivíduos.

Em Prathiba e Dhas (2013), são descritas técnicas baseadas na detecção de componentes dos corpos de pessoas como cabeça, ombros, braços e pernas ao invés de um único descritor de corpo inteiro. Esta técnica deve respeitar a distribuição geométrica e as proporções destes componentes em relação aos corpos. Este trabalho cita também modelos que realizam a segmentação da cena em áreas homogêneas, extraídas de vídeos que contem a movimentação de pedestres para a seleção e soma dos indivíduos.

Porém, estas estratégias baseadas em detecção de indivíduos e seus componentes tem se mostrado deficientes na medida em que a densidade das multidões aumenta. Nestes cenários, outras estratégias se fazem necessárias.

Kowcika e Sridhar (2016) discutem sobre a contagem de multidões em vídeos através de duas abordagens: baseadas em regiões de interesse (*Region of Interest* – ROI) que envolve o número estimado de pessoas em algumas regiões em determinada instância do tempo, e também na abordagem baseada na linha de interesse (*Line of Interest* – LOI) na qual a contagem é baseada no número de pessoas que cruzam esta linha.

Nos trabalhos de Sindagi e Patel (2018) e Luo, Lu e Zhang (2020), foram descritos os métodos pioneiros em contagem de multidões, com foco principal na evolução das técnicas baseadas em CNN e nos resultados alcançados em comparação com as outras técnicas, classificando dos métodos baseados em CNN com base na propriedade das redes

e na metodologia de inferência. Com base na propriedade das redes:

- CNNs básicas;
- Modelos com reconhecimento de escala;
- Modelos sensíveis ao contexto;
- Quadros multi-tarefa.

E com base na metodologia de inferência:

- Inferência baseada em patches;
- Inferência baseada na imagem inteira.

Os primeiros trabalhos que utilizaram CNN foram derivados de redes previamente treinadas como a VGG-16 (SIMONYAN; ZISSERMAN, 2014) e AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). A partir daí, várias arquiteturas de rede de aprendizado profundo tem sido adaptadas para a tarefa de processamento de imagens digitais, em especial para contagem de multidões.

Zhang et al. (2015) perceberam uma redução drástica de performance quando os métodos existentes eram aplicados a uma nova cena, diferente da base de dados utilizada para treinamento. Para resolver este problema, eles criaram uma rede que realiza o treinamento através de duas funções objetivas: contagem da multidão e estimativa de densidade.

Ciregan, Meier e Schmidhuber (2012) construíram uma rede de aprendizado profundo multicoluna para a tarefa de reconhecimento de imagens de multidões, alcançando uma considerável taxa de erro do modelo. Esta abordagem inspirou outros modelos que utilizam CNN para processamento de imagens digitais.

Em uma nova abordagem, Zhang et al. (2016) propuseram uma arquitetura de rede convolucional baseada em várias colunas *(Multi-Column CNN – MCNN)*, para imagens com perspectiva e densidade arbitrárias. Esta rede é composta por três tamanhos diferentes de *kernel* (pequeno, médio e grande), para atender a diferentes escalas de tamanho de objetos presentes na rede.

Na abordagem de Boominathan, Kruthiventi e Babu (2016), foram combinadas redes convolucionais profundas e rasas *(deep and shallow)* para prever o mapa de densidade de uma imagem de multidão, na tentativa de capturar mais informações semânticas destas imagens.

Na tentativa de obter melhor performance na contagem de multidões, Zeng et al. (2017), desenvolveram uma arquitetura de rede de coluna única multiescala *(Multi-scale CNN – MSCNN)*, que é ao mesmo tempo precisa e econômica para aplicações práticas, pois é capaz de gerar recursos relevantes à escala para desempenhos mais altos de contagem de multidões.

Baseado na ideia de que o melhor desempenho pode ser obtido pelo treinamento de regressores e um pequeno conjunto de correções nestes dados, Sam, Surya e Babu (2017) propuseram uma CNN de comutação que seleciona um regressor ideal adequado para um determinado *patch* de entrada. Esta rede é similar a uma rede multi-coluna, mas com diferentes campos receptivos e o uso de um classificador.

Um dos grandes desafios na contagem de multidões envolve as drásticas mudanças de escalas e perspectivas nas imagens. As abordagens que lidam com filtros e estimativas baseadas em *patchs* são, em geral, computacionalmente caras.

Diante deste desafio, Shi et al. (2018) propuseram uma CNN com reconhecimento de perspectiva *(Perspective-aware CNN – PACNN)*, para prever diretamente um mapa de perspectiva na rede e usa uma camada de ponderação para combinar de forma adaptável as saídas de densidade dos mapas de recursos de várias escalas.

Outra rede baseada em perspectiva, neste caso espacial, foi desenvolvida por Liu et al. (2018), onde uma rede composta de módulos é usada para a contagem de multidões. Nela, as características globais da imagem são extraídas e em seguida usadas para estimar um mapa inicial de densidade. Em seguida, um módulo é aplicado para localizar iterativamente regiões da imagem com um mecanismo baseado em um transformador espacial para refinar a região do mapa de densidade assistida por um aprendizado residual.

Li, Zhang e Chen (2018) desenvolveram uma CNN mais profunda com convoluções dilatadas que alcançou um desempenho de ponta, chamada *Congested Scene Recognition Network* (CSRNet). Ela amplia os campos receptivos empilhando convoluções dilatadas. Este trabalho serviu de inspiração e base para novos modelos aplicados na contagem de multidões.

Um destes trabalhos utilizou parte da CSRNet como *front-end*, espinha dorsal ou *backbone* da rede, empilhando novas camadas na tentativa de alcançar melhores resultados e redução na taxa de erro (Yan et al., 2019). Inclusive, duas das propostas deste trabalho foram inspiradas em ideias extraídas desta abordagem que serviu também como base para análise e comparação dos resultados.

Outro trabalho já mencionado anteriormente de um modelo de rede chamado PACNN (SHI et al., 2018) propôs uma revisão alterando o *backbone* da rede, inicialmente com parte a rede VGG-16, pela rede CSRNet previamente treinada. Estas alterações

permitiram que o modelo pudesse alcançar melhores resultados e ficou conhecida como PACNN + CSRNet.

No entanto, a rede CSRNet fornece um campo receptivo fixo para diferentes escalas de pessoas, permanecendo vulnerável às escalas intracenas altamente variantes, por isso tem um bom desempenho em escalas intermediárias, mas se comporta de maneira relativamente ruim em escalas mais amplas.

De qualquer forma, analisando a literatura, os modelos que exploram a transferência de aprendizado tem sido aqueles mais utilizados por proporcionarem, na maioria dos casos, um menor esforço na etapa de treinamento. Além da rede VGG-16, outras redes previamente treinadas também são utilizadas na tarefa de contagem de multidões.

Um exemplo é a rede DUBNet (OH; OLSEN; RAMAMURTHY, 2020), que optou por utilizar a rede ResNet-50 (He et al., 2016) como *backbone* e justificou a escolha pelo fato de que apesar desta rede ser muito mais profunda que a VGG-16, a ResNet-50 (3,8B FLOPs) tem uma complexidade muito menor do que a VGG-16 (15,3B FLOPs).

Um trabalho recente (Sindagi; Patel, 2020) propôs uma rede hierárquica para contagem de multidão baseada na atenção chamada de HA-CCN (*Hierarchical Attention-based Crowd Counting Network*), que emprega mecanismos de atenção em vários níveis para aprimorar seletivamente os recursos da rede e que também explora a transferência de aprendizado, utilizando os pesos da rede VGG-16.

Estes modelos que exploram a transferência de aprendizado adaptam uma rede ou parte de uma rede previamente treinada, empilhando novas camadas que serão treinadas ou retreinadas para a solução de novos problemas em diferentes domínios.

Neste trabalho, propomos o uso de redes neurais artificiais profundas, inspiradas em algumas destas abordagens e também em outras novas abordagens, na tentativa de superar o estado da arte na tarefa da estimativa de contagem de multidões.

3 PROPOSTAS DE ARQUITETURAS DE REDES NEURAIS ARTIFICIAIS PROFUNDAS PARA A SOLUÇÃO DO PROBLEMA DA ESTIMATIVA DE CONTAGEM DE MULTIDÕES

Conforme mencionado nos capítulos anteriores, as redes neurais artificiais profundas têm sido amplamente utilizadas em tarefas da área visão computacional e na solução do problema da contagem de multidões.

Na literatura foram encontradas diferentes arquiteturas de rede, sendo algumas delas profundas do ponto de vista do número de camadas e outras mais rasas. Existem também aquelas com múltiplos fluxos com tamanhos de *kernel* variados, dentre outras.

Algumas limitações apontadas nas CNNs tradicionais somadas aos desafios crescentes na área da visão computacional, motivaram também o surgimento de novas abordagens, como é o caso das redes neurais artificiais em cápsulas.

Neste Capítulo, serão descritas as propostas de redes neurais artificias profundas desenvolvidas para solucionar o problema da contagem de multidões.

3.1 Arquitetura Ca-CCNet

Uma das hipóteses é a de que o uso de uma CapsNet possa ser mais eficiente ao lidar com cenas de multidões, levando em consideração também a quantidade limitada de imagens de algumas das bases utilizadas neste tipo de tarefa. Uma CapsNet, teoricamente, precisa de uma quantidade menor de dados para aprender a representação das entidades analisadas.

Além disso, uma CapsNet lida melhor com as limitações encontradas nas CNNs, devido as perdas causadas pela operação de agrupamento e a incapacidade de entender a relação espacial entre as *features* (SABOUR; FROSST; HINTON, 2017).

Esta primeira proposta, chamada de *CapsNet for Crowd Counting Network* (Ca-CCNet), foi inspirada em uma arquitetura de rede neural em cápsula chamada de SegCaps, desenvolvida por LaLonde e Bagci (2018), para a solução de um problema de segmentação de imagens médicas. Segundo os autores, foi o primeiro uso de uma arquitetura CapsNet para segmentação de objetos na literatura.

As modificações propostas pelos autores, principalmente no algoritmo de roteamento dinâmico e no compartilhamento das matrizes de transformação, possibilitaram a redução da carga de memória e dos parâmetros em relação a implementação da cápsula

original, permitindo também a utilização de imagens com maior resolução, enquanto as primeiras CapsNets que surgiram eram restritas a imagens de entrada muito pequenas.

A arquitetura da Ca-CCNet, apresentada na Figura 17 e detalhada na Tabela 1, consiste basicamente em: camadas de convolução e camadas de cápsula. As duas primeiras camadas convolucionais regulares da rede tentam extrair *features* de baixo nível.

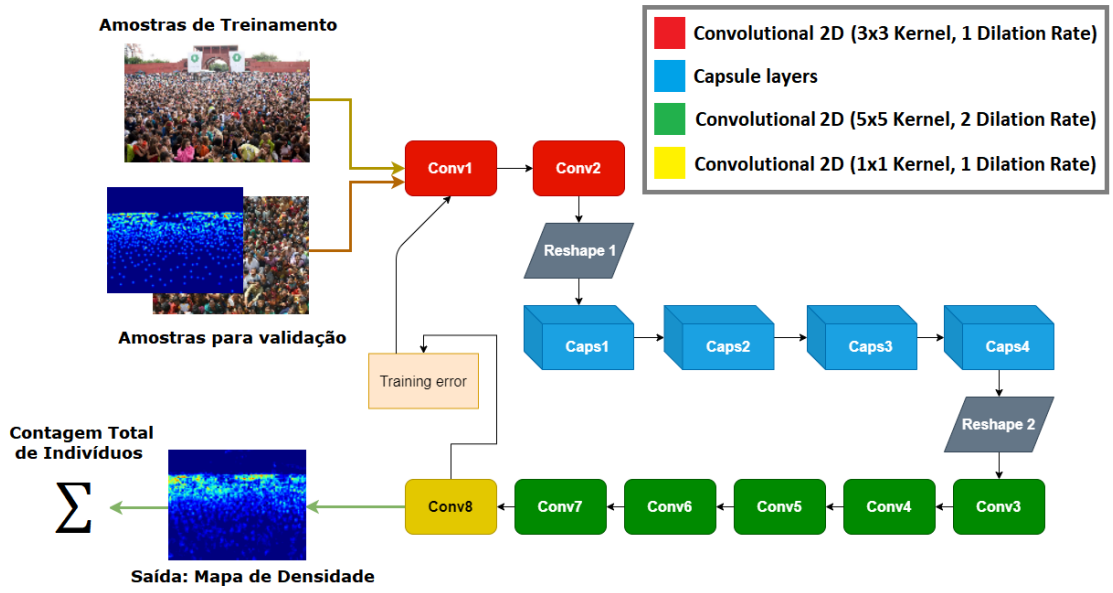
As saídas serão remodeladas e convertidas em vetores de ativação nas próximas camadas de cápsulas. Esta proposta trabalha com mapas com 1/4 do seu tamanho original, devido a configuração da arquitetura proposta.

A dimensão das cápsulas é definida neste trabalho, para todas as propostas, como **átomos** e o *padding* utilizado foi o SAME, também para todas as propostas, e significa que o tamanho dos mapas de recursos de saída são os mesmos que os mapas de recursos de entrada.

Em relação a proposta original, removemos as camadas finais utilizadas originalmente para segmentação e reconstrução por camadas convolucionais dilatadas com janela de tamanho fixo igual a 5×5 e fator de dilatação igual a 2. A escolha desta estrutura final da arquitetura foi inspirada na rede CSRNet (Li; Zhang; Chen, 2018).

O uso de convoluções dilatadas oferece um campo de visão mais amplo ao mesmo custo computacional, definindo um espaçamento entre os valores em um *kernel* conforme

Figura 17 – Primeira proposta de rede neural para a solução do problema da contagem de multidões chamada de Ca-CCNet. Destaque para as camadas intermediárias em cápsulas na cor azul.



Fonte: Elaborada pelo autor

Tabela 1 – Configuração da Ca-CCNet. As camadas convolucionais são denotadas por "Conv2D(Filtros × *Kernel* × *Strides* × Taxa de dilatação)" e das camadas em cápsulas "Caps(*Kernel* × Número de Cápsulas × Átomos × *Strides* × Número de roteamentos)".

Mapas de recursos	Nome
Input: (256x256x3)	inputlayer
Conv2D(64x3x1x1)	block1_conv1
Conv2D(64x3x1x1)	block1_conv2
Reshape	Res01
Caps(5x2x32x2x1)	Caps01
Caps(5x2x64x1x3)	Caps02
Caps(5x2x128x2x1)	Caps03
Caps(1x1x256x1x3)	Caps04
Reshape	Res02
Conv2D(256x5x1x2)	block2_conv1
Conv2D(256x5x1x2)	block2_conv2
Conv2D(256x5x1x2)	block2_conv3
Conv2D(128x5x1x2)	block2_conv4
Conv2D(64x5x1x2)	block2_conv5
Conv2D(1x1x1x1)	y_out

Fonte: Elaborada pelo autor

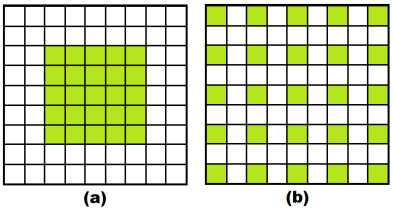
ilustra a Figura 18. Por exemplo, um *kernel* 5×5 com uma taxa de dilatação de 2 terá o mesmo campo de visão que um *kernel* 9×9, enquanto usa menos parâmetros.

Durante o treinamento, a rede aprenderá gradualmente uma matriz de transformação entre pares de cápsulas para definição do relacionamento parte-todo das entidades analisadas.

Este grupo de neurônios chamados cápsulas permite o armazenamento de informações em vetores ao invés de valores escalares, na tentativa de modelar melhor as relações hierárquicas dentro da representação interna do conhecimento de uma rede neural artificial profunda.

Levando em consideração os desafios que envolvem uma tarefa de contagem de multidões, principalmente quando o número de amostras para treinamento é reduzido e não representam toda a complexidade destes cenários, esta arquitetura foi proposta na tentativa de melhor aprender a representação, sem ampliar e aprofundar a rede.

Figura 18 – Exemplo de dilatação com janela de 5×5 com taxa de dilatação igual a 1(a); e com taxa de dilatação igual a 2(b).



Fonte: Elaborada pelo autor

Até onde foi possível verificar na literatura da área, esta foi a primeira proposta de uso de uma CapsNet para a solução do problema de contagem de multidões.

3.2 Arquitetura QU-CCNet

Uma arquitetura de rede que inspirou nossa segunda proposta foi a U-Net, desenvolvida por Ronneberger, Fischer e Brox (2015), para segmentação semântica de imagens biomédicas. É uma CNN composta por um caminho de codificação usado para capturar o contexto e outro caminho de decodificação usado para definir a precisão da localização.

A segmentação de imagens digitais visa o particionamento de uma imagem em regiões disjuntas com algum significado, de acordo com a aplicação. Neste caso, cada pixel da imagem analisada deverá pertencer a uma região.

Uma arquitetura U-Net oferece um tipo de previsão densa que tem como saída uma imagem de alta resolução, onde cada pixel é associado a uma classe correspondente. Ela tem este nome pelo formato em "U" da rede.

Uma hipótese deste trabalho é a de que este tipo de arquitetura de rede pode alcançar melhores resultados pela combinação de diferentes camadas da rede, utilizadas tanto para codificação quanto para decodificação, na tentativa de lidar melhor com a perda da informação da localização das *features* ao longo da rede.

Esta segunda proposta é chamada de *Quasi U-Net for Crowd Counting Network* (QU-CCNet), desenvolvida para a solução da tarefa de contagem de multidões e inspirada na arquitetura U-Net.

A proposta desta arquitetura está representada na Figura 19 e tem este nome pelo fato de se trabalhar com mapas de densidade com 1/4 do tamanho original da imagem, por isso foram removidas duas camadas decodificadoras nesta proposta que se utilizadas formariam a arquitetura em formato de "U" completa, por isso a denominação "quase em U" (*Quasi* U-Net).

A rede é composta por camadas de convolução, de agrupamento e camadas de expansão com convoluções transpostas. Usamos também camadas de *Dropout* que são um tipo comum de camada de regularização usada para melhorar o desempenho da rede e minimizar *overfitting*.

O *backbone* da rede é composto pelas 13 camadas convolucionais importadas da rede VGG-16 (SIMONYAN; ZISSERMAN, 2014), explorando a transferência de aprendizado e o compartilhamento de pesos, visando também a redução na etapa de treinamento. Em seguida, utilizamos mais duas camadas de convolução seguidas de uma primeira camada de convolução transposta. A configuração da rede está disposta na Tabela 2.

A partir do momento que avançamos na expansão do decodificador, percebemos a necessidade da utilização de camadas mais inteligentes, por isso as camadas de *upsampling*, que geralmente são utilizadas em redes do tipo U-Net, foram substituídas por camadas de convolução transposta.

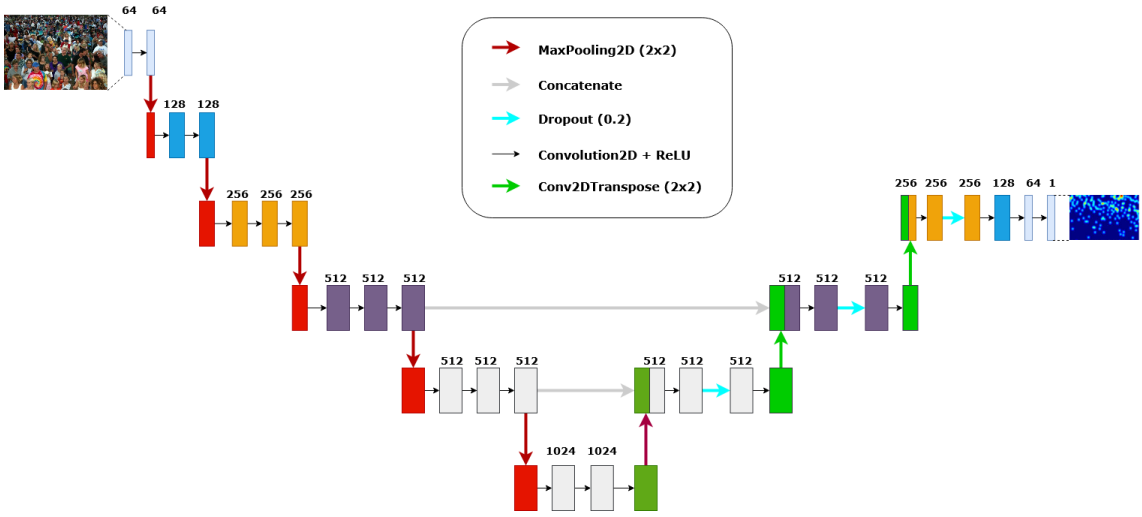
Na operação de convolução transposta, o *kernel* é aprendido (assim como na operação de convolução) durante o treinamento, ao contrário de outras operações que utilizam apenas uma ampliação simples da imagem usando o vizinho mais próximo.

Um dos fatores que também inspirou esta segunda proposta está relacionado ao uso de um decodificador na rede (lado direito da rede ou caminho de expansão), que tenta recuperar a informação perdida para melhorar a precisão da localização, consequentemente os resultados.

Em cenários complexos como são as cenas de multidões, melhorar a precisão da localização pode ser um fator importante que pode contribuir significativamente com o sucesso na tarefa de estimativa de contagem.

Com o sucesso alcançado nas tarefas de segmentação, espera-se com esta abordagem superar alguns dos desafios encontrados na tarefa de contagem de multidões.

Figura 19 – Proposta de arquitetura de rede neural chamada QU-CCNet, desenvolvida para a solução do problema de contagem de multidões. Cada seta corresponde a uma operação, de acordo com a legenda. O número de canais é indicado na parte superior de cada caixa que representa uma camada de convolução com *kernel* de tamanho fixo igual a 3×3.



Fonte: Elaborada pelo autor

Tabela 2 – Configuração da QU-CCNet. As camadas convolucionais tradicionais e transpostas são denotadas por "Conv2D(Filtros × *Kernel* × *Strides* × Taxa de dilatação)".

Mapas de recursos	Nome
Input: (512x512x3)	inputlayer
Conv2D(64x3x1x1)	block1_conv1
Conv2D(64x3x1x1)	block1_conv2
MaxPooling2D(2x2)	block1_pool1
Conv2D(128x3x1x1)	block2_conv1
Conv2D(128x3x1x1)	block2_conv2
MaxPooling2D(2x2)	block2_pool2
Conv2D(256x3x1x1)	block3_conv1
Conv2D(256x3x1x1)	block3_conv2
Conv2D(256x3x1x1)	block3_conv3
MaxPooling2D(2x2)	block3_pool3
Conv2D(512x3x1x1)	block4_conv1
Conv2D(512x3x1x1)	block4_conv2
Conv2D(512x3x1x1)	block4_conv3
MaxPooling2D(2x2)	block4_pool4
Conv2D(512x3x1x1)	block5_conv1
Conv2D(512x3x1x1)	block5_conv2
Conv2D(512x3x1x1)	block5_conv3
MaxPooling2D(2x2)	block5_pool5
Conv2D(1024x3x1x1)	block6_conv1
Conv2D(1024x3x1x1)	block6_conv2
Conv2DTranspose(512x3x2x1)	ublock7_conv1
Concatenate: ublock7_conv1 + block5_conv3	ublock7_concat1
Conv2D(512x3x1x1)	ublock7_conv2
Dropout(0.2)	ublock7_drop1
Conv2D(512x3x1x1)	ublock7_conv3
Conv2DTranspose(512x3x2x1)	ublock8_conv1
Concatenate: ublock8_conv1 + block4_conv3	ublock8_concat2
Conv2D(512x3x1x1)	ublock8_conv2
Dropout(0.2)	ublock8_drop2
Conv2D(512x3x1x1)	ublock8_conv3
Conv2DTranspose(256x3x2x1)	ublock9_conv1
Concatenate: ublock9_conv1 + block3_conv3	ublock9_concat3
Conv2D(256x3x1x1)	ublock9_conv2
Dropout(0.2)	ublock9_drop3
Conv2D(256x3x1x1)	ublock9_conv3
Conv2D(128x3x1x1)	block10_conv1
Conv2D(64x3x1x1)	block10_conv2
Conv2D(1x1x1x1)	y_out

Fonte: Elaborada pelo autor

3.3 Arquitetura CaTL-CCNet

A transferência de aprendizado permite que um conhecimento adquirido ao resolver determinado problema possa ser aplicado para resolver outro problema, respeitando um conjunto de premissas intrínsecas. Um modelo que foi previamente treinado em alguma tarefa é reutilizado para a solução de uma nova tarefa.

Treinar a rede "do zero" pode ser um procedimento caro e demorado. Por isso,

esta estratégia de transferência de aprendizado tem sido amplamente utilizada em tarefas recentes de visão computacional pelos benefícios que ela proporciona, como a possibilidade da redução dos custos na fase de treinamento (principalmente tempo e processamento), além da utilização de um tipo de aprendizado na solução de um novo problema.

Um modelo previamente treinado pode, quando permitido, ser adaptado em sua estrutura. Uma das estratégias mais comuns envolve a remoção de parte da rede e a adição de novas camadas para adaptar à nova rede ao domínio do problema e evitar ajustes excessivos.

É possível também "congelar" uma parte da rede e treinar somente outra parte, principalmente se o novo conjunto de dados for muito pequeno ou se a rede for complexa ou pesada (uma rede grande do ponto de vista de parâmetros e camadas), portanto difícil de treinar.

Neste caso, como as camadas iniciais já foram treinadas e aprenderam a reconhecer *features* de nível baixo, as camadas finais serão treinadas para combinar estas características em favor do aprendizado de novas entidades.

Como vimos nos capítulos anteriores, as CNNs podem perder informações ao logo da rede e necessitam de uma quantidade significativa de amostras de treinamento ou de ampliação da arquitetura para aprender a representação de um objeto em diferentes perspectivas. Por este motivo também foram propostos modelos baseados em CapsNets.

No entanto, apesar desta perda, algumas redes têm alcançados bons resultados e ainda estão disponíveis para reutilização.

Nesta terceira proposta chamada de *CapsNet with Transfer Learning for Crowd Counting Network* (CaTL-CCNet), realizou-se a combinação de parte de uma CNN previamente treinada, com camadas em capsulas.

A hipótese aqui é a de que uma rede previamente treinada tem alta capacidade de reconhecimento de *features* de baixo nível (curvas, setas, cores, etc.), conforme exemplificado na Figura 20, e que camadas em cápsula no final da rede podem combinar e processar estas informações para aprender melhor a representação das entidades e objetos, apesar das camadas de *pooling* presentes.

Entende-se que um número reduzido destas camadas de agrupamento não é tão prejudicial se combinadas com uma estratégia que explora ou combina de uma melhor maneira os recursos aprendidos nas camadas anteriores da rede.

Esta arquitetura tenta simular um caminho de processamento visual construtivo, em que a representação vai sendo construída ao longo da rede, na qual as primeiras camadas da rede (CNN) simulam o processamento visual de nível inferior e intermediário

e que as camadas finais (CapsNet) simulam o processamento visual de nível superior.

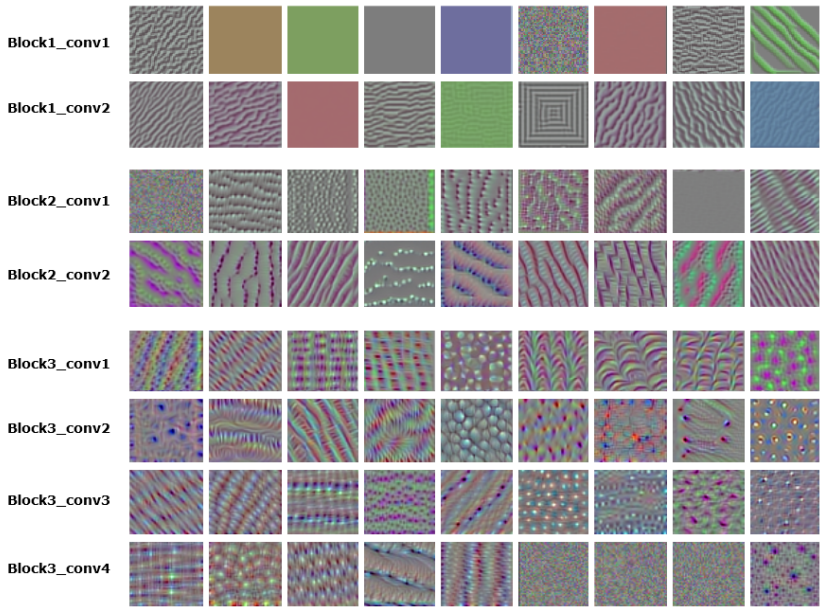
As características extraídas das primeiras camadas serão combinadas, permitindo distinguir quais estímulos pertencem ao objeto de análise, no nosso caso a multidão, e quais pertencem a outros como o plano de fundo.

A arquitetura da rede é composta pelas 12 primeiras camadas convolucionais da rede VGG-19 (SIMONYAN; ZISSERMAN, 2014), combinadas com camadas em cápsula para a produção dos mapas de densidade. As cápsulas serão então utilizadas para combinar estas características previamente treinadas visando um melhor resultado. A proposta está descrita na Tabela 3 e ilustrada na Figura 21.

Um trabalho que também inspirou esta proposta foi o CSRNet (Li; Zhang; Chen, 2018), que é caracterizado por uma CNN que utiliza em sua estrutura parte da VGG-16 (SIMONYAN; ZISSERMAN, 2014) combinada com camadas convolucionais dilatadas no final da rede.

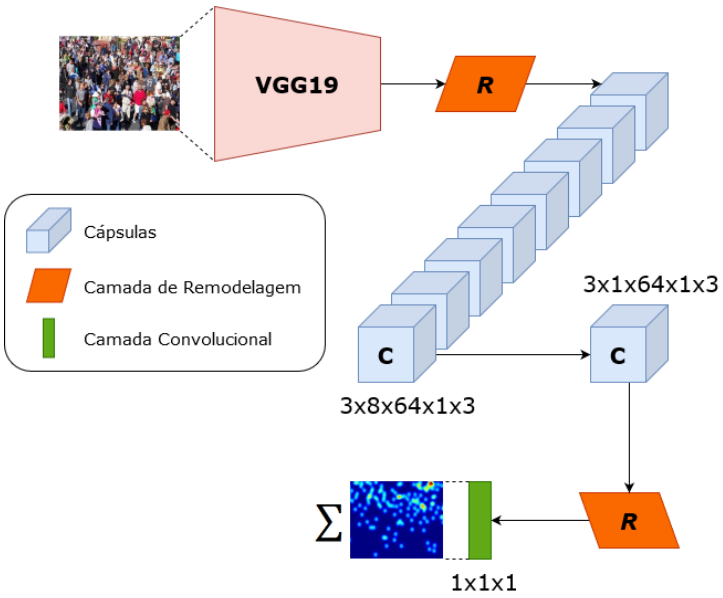
A CSRNet possui arquitetura semelhante à da nossa proposta CaTL-CCNet, mas diferente em dois aspectos principais: (i) CSRNet utiliza parte da VGG-16 enquanto a proposta CaTL-CCNet utiliza parte da VGG-19; e (ii) foram substituidas as convoluções dilatadas presentes na CSRNet por camadas em cápsulas. Na Tabela 4 temos a comparação destas duas abordagens.

Figura 20 – Características de baixo nível e nível intermediário extraídas das primeiras camadas da nossa terceira proposta de rede chamada CaTL-CCNet, importadas da rede VGG-19. Cada linha representa alguns canais (ou mapas) de uma dada camada convolucional.



Fonte: Elaborada pelo autor

Figura 21 – Proposta de arquitetura de rede neural chamada CaTL-CCNet, desenvolvida para a solução do problema de contagem de multidões. As camadas em cápsulas estão configuradas da seguinte forma: "Caps(*Kernel* × Cápsulas × Átomos × *Strides* × Roteamentos)".



Fonte: Elaborada pelo autor

Tabela 3 – Configuração da CaTL-CCNet. As camadas convolucionais são denotadas por "Conv2D(Filtros × *Kernel* × *Strides* × Taxa de dilatação)" e das camadas em cápsulas "Caps(*Kernel* × Número de Cápsulas × Átomos × *Strides* × Número de roteamentos)".

Mapas de recursos	Nome
Input:(512x512x3)	inputlayer
Conv2D(64x3x1x1)	block1_conv1
Conv2D(64x3x1x1)	block1_conv2
MaxPooling2D(2x2)	block1_pool1
Conv2D(128x3x1x1)	block2_conv1
Conv2D(128x3x1x1)	block2_conv2
MaxPooling2D(2x2)	block2_pool2
Conv2D(256x3x1x1)	block3_conv1
Conv2D(256x3x1x1)	block3_conv2
Conv2D(256x3x1x1)	block3_conv3
Conv2D(256x3x1x1)	block3_conv4
MaxPooling2D(2x2)	block3_pool3
Conv2D(512x3x1x1)	block4_conv1
Conv2D(512x3x1x1)	block4_conv2
Conv2D(512x3x1x1)	block4_conv3
Conv2D(512x3x1x1)	block4_conv4
Reshape	Res01
Caps(3x8x64x1x3)	Caps01
Caps(3x1x64x1x3)	Caps02
Reshape	Res02
Conv2D(1x1x1x1)	Layer_Out

Fonte: Elaborada pelo autor

O que se deseja mostrar com esta proposta é que o uso de uma estratégia baseada em transferência de aprendizado combinada com camadas em cápsulas pode ser tão ou mais eficiente quanto outras propostas baseadas puramente em CNN.

Tabela 4 – Comparação entre CaTL-CCNet e CSRNet. As camadas convolucionais são denotadas por "Conv2D-(Filtros)-(Taxa de dilatação)" e das camadas em cápsulas "Caps(*Kernel* × Número de Cápsulas × Átomos × *Strides* × Número de roteamentos)". A arquitetura da CSRNet contém duas camadas a mais (desconsiderando o Reshape), destacadas em vermelho, do que a nossa proposta CaTL-CCNet.

CaTL-CCNet	**CSRNet**
Conv2D-64-1	Conv2D-64-1
Conv2D-64-1	Conv2D-64-1
MaxPooling2D	
Conv2D-128-1	Conv2D-128-1
Conv2D-128-1	Conv2D-128-1
MaxPooling2D	
Conv2D-256-1	Conv2D-256-1
Conv2D-256-1	Conv2D-256-1
Conv2D-256-1	Conv2D-256-1
Conv2D-256-1	
MaxPooling2D	
Conv2D-512-1	Conv2D-512-1
Conv2D-512-1	Conv2D-512-1
Conv2D-512-1	Conv2D-512-1
Conv2D-512-1	
Reshape	
Caps(3x8x64x1x3)	Conv2D-512-2
Caps(3x1x64x1x3)	Conv2D-512-2
	Conv2D-512-2
	Conv2D-256-2
	Conv2D-128-2
	Conv2D-64-2
Reshape	
Conv2D(1x1x1x1)	Conv2D(1x1x1x1)

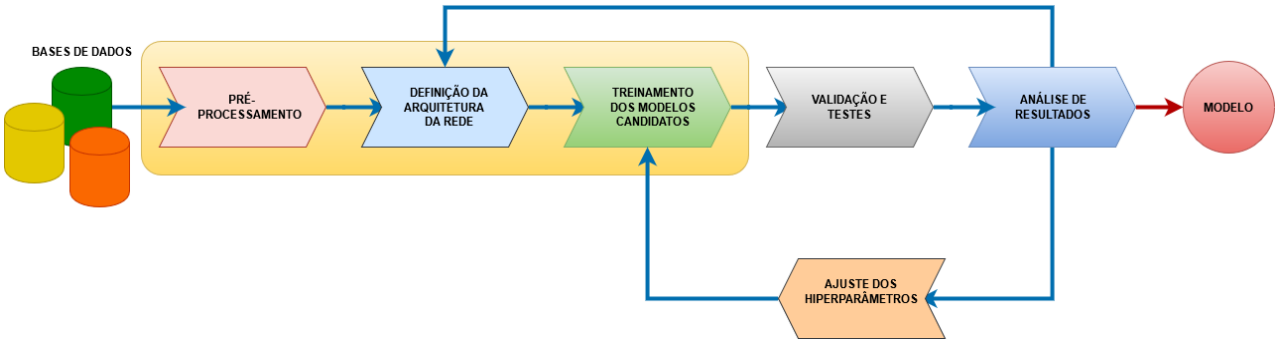
Fonte: Elaborada pelo autor

4 EXPERIMENTOS E ANÁLISE DOS RESULTADOS

Neste Capítulo serão apresentados os resultados obtidos nos experimentos realizados em cada uma das abordagens propostas. Em geral, todas as propostas passaram pelo mesmo processo de treinamento e testes.

Sendo assim, a metodologia utilizada em todas as propostas foi a mesma e está ilustrada na Figura 22.

Figura 22 – Metodologia utilizada para o desenvolvimento e escolha dos modelos propostos neste trabalho.



Fonte: Elaborada pelo autor

Este Capítulo está organizado da seguinte maneira. Na Seção 4.1 é descrita toda a etapa de treinamento, atividades e recursos associados. Na Seção 4.2 é explicada a etapa de testes. Na Seção 4.3 serão apresentados os resultados obtidos, qualitativos e quantitativos, por cada uma das propostas. Por fim, na Seção 4.4 serão comparados todos os resultados alcançados por todas as propostas.

4.1 Etapa de treinamento

Foi utilizado o mesmo processo de treinamento em todas as propostas deste trabalho, ou seja, as mesmas amostras de treinamento e testes previamente preparadas e os mesmos parâmetros iniciais e métricas para análise e comparação dos resultados.

Em geral, este processo consiste de atividades como a seleção e preparação das bases de dados, da definição e desenvolvimento de arquiteturas das redes neurais artificiais profundas, análise dos resultados, primeiramente com amostras reduzidas para validar o modelo e em seguida com o conjunto completo, e ajuste dos hiperparâmetros.

Nas próximas seções serão detalhadas as atividades e recursos envolvidos na etapa de treinamento.

4.1.1 Bases de dados

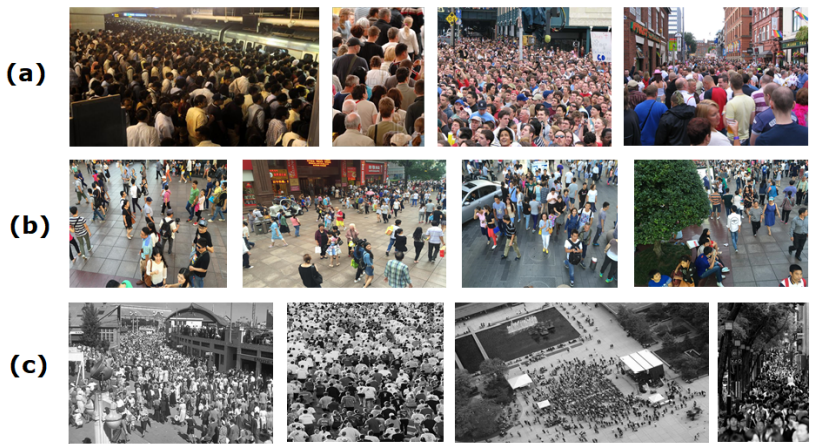
Foram realizados experimentos das nossas propostas em dois conjuntos de dados: UCF_CC_50 (IDREES et al., 2013) e ShanghaiTech (ZHANG et al., 2016). Amostras de imagens extraídas destes repositórios estão ilustradas na Figura 23.

A base ShanghaiTech é um conjunto de 1198 imagens de multidões de média e alta densidade, dividido em dois conjuntos: A e B, contendo as coordenadas das 330.165 pessoas, no qual:

- **Parte A**: 482 imagens de multidões mais densas recuperadas da Internet, sendo 300 delas utilizadas para treinamento e 182 para testes;
- **Parte B**: 716 imagens de multidões um pouco mais esparsas da região metropolitana de Xangai/China, sendo 400 delas separadas para treinamento e 316 para testes.

UCF_CC_50 é uma base de imagens da University of Central Florida (Center for Research in Computer Vision - UCF), que possui imagens de multidões densas coletadas manualmente do FLICKR. Consiste de um pequeno conjunto que contém 50 imagens em escala de cinza, com as coordenadas de um total de 63.974 pessoas. A limitação do conjunto em relação quantidade de imagens e sua característica (monocromáticas) pode exigir maior esforço no treinamento.

Figura 23 – Amostras de imagens recolhidas dos conjuntos de dados. As linhas correspondem aos conjuntos ShanghaiTech Parte A (a); ShanghaiTech Parte B (b); e UCF_CC_50 (c).



Fonte: Elaborada pelo autor

4.1.2 Mapas de densidade

As imagens contidas nas bases utilizadas neste trabalho possuem as anotações das coordenadas no plano da cabeça de cada indivíduo identificado. Com estas anotações, serão gerados os mapas de densidade utilizados para treinar o modelo, conforme ilustra a Figura 24.

Neste trabalho utilizamos abordagens tradicionais disponíveis na literatura para geração dos mapas de densidade (Li; Zhang; Chen, 2018; ZHANG et al., 2016). A função utilizada para geração de mapas de densidade está representada pela Equação 4.1.

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x), \, com \; \sigma_i = \beta \bar{d}_i \tag{4.1}$$

Para cada objeto alvo x_i no *ground-truth* δ, usa-se $\bar{d}_i$ para indicar a distância média dos k vizinhos mais próximos. Seguindo a configuração escrita por Zhang et al. (2016), para desfocar todas as anotações fixou-se $\beta = 0,3$ e $k = 2$.

O mapa é então obtido pela computação da convolução $\delta(x - x_i)$ por um *kernel* gaussiano com desvio patrão parametrizado G_{σ_i}, onde x corresponde a posição da anotação na imagem. Neste trabalho, para a geração dos mapas, foram adotados valores fixos de σ, conforme ilustra a Tabela 5, baseado no tamanho médio das cabeças.

Tabela 5 – Valor de desvio padrão utilizado na geração dos mapas de densidade de treinamento.

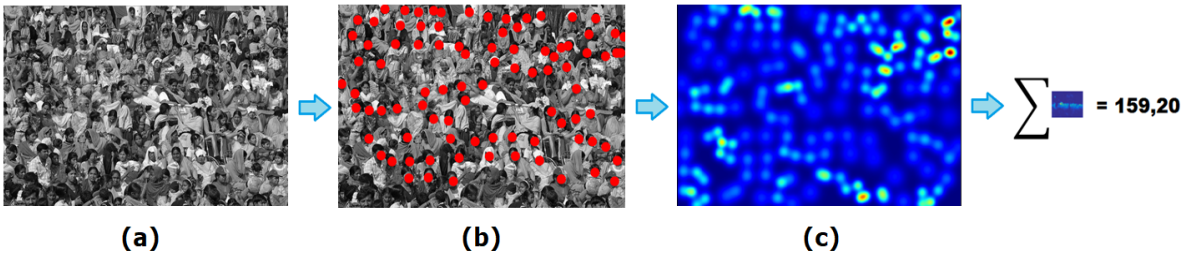
Repositórios	Desvio Padrão (σ)
ShanghaiTech Part A	5
ShanghaiTech Part B	15
UCF_CC_50	4

Fonte: Elaborada pelo autor

Os mapas são construídos ao desfocar cada anotação de cabeça na imagem usando filtro gaussiano normalizado para 1, considerando a distribuição espacial nas imagens a partir de cada conjunto de dados, conforme ilustra a Figura 25.

As cores vermelho e azul no mapa correspondem a regiões de alta e baixa densidade, respectivamente. Cores intermediárias (por exemplo: amarelo e laranja) representam variações desta densidade. Na Figura 26 temos exemplos de variações destas distribuições de cores nos mapas de densidade.

Figura 24 – As anotações contidas nas bases utilizadas na tarefa de contagem de multidões. Imagem original (a); anotações que correspondem as coordenadas das cabeças das pessoas identificadas na imagem (cor vermelha) (b); e o mapa de densidade criado (c). Por fim, este mapa é então integrado para estimar o número total de pessoas.



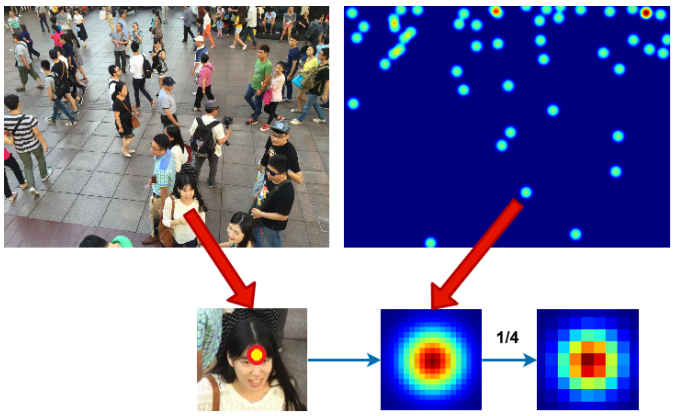
Fonte: Elaborada pelo autor

4.1.3 Ajuste de parâmetros

Uma vez definidas as propostas de arquiteturas de rede, é necessário determinar um protocolo de treinamento. Apesar de serem utilizadas propostas de redes distintas (CapsNet e CNN), o protocolo de treinamento foi mantido para ambas as abordagens. Trata-se de treinamento supervisionado, respeitando a característica da tarefa em que os alvos são os mapas de densidade.

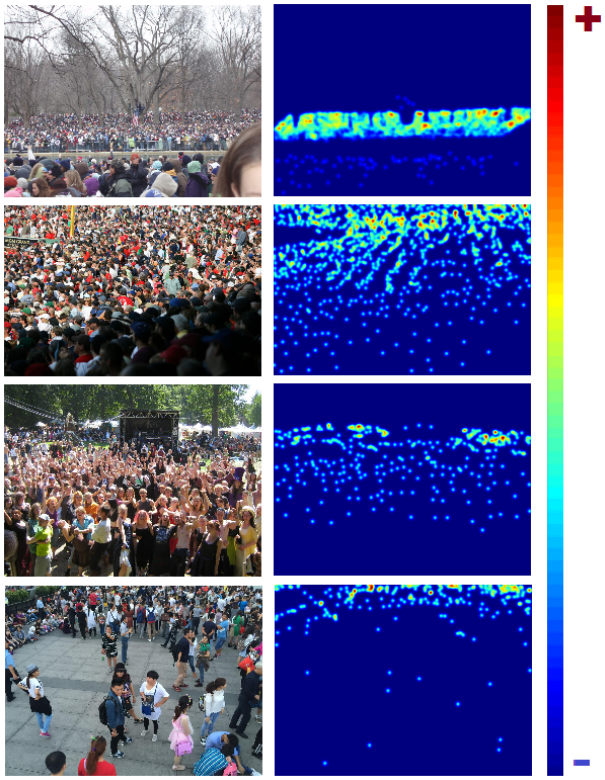
A primeira etapa diz respeito a preparação das amostras de treinamento, ou pré-processamento. Esta etapa prévia visa adaptar os dados para melhor manipulação pela rede ou otimizá-los para um melhor desempenho. Estas amostras serão utilizadas para otimizar os pesos do modelo durante a fase de treinamento.

Figura 25 – Filtro gaussiano para a construção do mapa de densidade. Cada posição da cabeça é convertida em uma distribuição gaussiana (distribuições sobrepostas são somadas).



Fonte: Elaborada pelo autor

Figura 26 – Mapa de densidade criado baseado nas anotações. As cores vermelho e azul representam, respectivamente, regiões de maior e menor densidade.



Fonte: Elaborada pelo autor

Para a etapa de treinamento foram realizados ajustes no tamanho das imagens, principalmente em conjuntos com grandes variações de tamanho (como a ShanghaiTech Parte A). Estes ajustes se referem a ampliação e corte para que fosse possível extrair delas amostras (*patches*) de tamanho fixo, pois o redimensionamento das imagens poderia causar distorção e comprometer o resultado. Este ajuste corresponde ao preenchimento da área ausente com valores zero e destas imagens foram extraídos *patches* com novas resoluções (256×256 e 512×512).

Além disso, as imagens de entrada foram transformadas por uma técnica comum de pré-processamento que envolve subtrair a imagem média para tornar os dados de entrada centralizados em zero. Assim, é esperado um melhor desempenho, pois as funções de ativação (ReLU) acabam sendo mais responsivas às alterações de peso em torno de zero.

Nossas propostas foram desenvolvidas em *Python*, utilizando as bibliotecas Keras / Tensorflow (ABADI et al., 2016). O otimizador Adam (KINGMA; BA, 2015) foi usado para treinamento dos modelos com uma taxa de aprendizado de $1,0 \times 10^{-5}$. Foi utilizado o método de inicialização de pesos *He Normal* (He et al., 2015), que combinado com ativação ReLU, tenta evitar o problema de *gradient vanishing*. Outra forma utilizada nas nossas propostas foi por meio de inicialização gaussiana com desvio padrão de $0,01$.

A função de perda utilizada foi baseada no erro médio quadrado L_{MSE} (*Mean Squared Error Loss*), definida pela Equação 4.2:

$$L_{MSE} = \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} |C_i - C_i^T|^2 \tag{4.2}$$

na qual N_{tr} corresponde ao total de amostras de treinamento, C_i é o número total de contagem predito e C_i^T corresponde a contagem total verdadeira.

As imagens foram testadas em cada uma das redes propostas neste trabalho. Amostras de testes são utilizadas também para validar o poder de generalização dos modelos. Cada repositório de imagens ShanghaiTech Parte A e Parte B já possuem um subconjunto de imagens para testes.

Não existe um subconjunto de testes previamente definido para o repositório de imagens UCF_CC_50. Neste caso, foi realizada a validação cruzada em 5 dobras *(5-fold cross-validation)* na qual o conjunto foi dividido aleatoriamente em 5 subconjuntos de 10 imagens cada. A partir destes subconjuntos foi avaliada a capacidade de generalização do modelo por meio da média dos resultados obtidos usando-se cada subconjunto para teste enquanto os demais forneciam dados para o treinamento.

4.1.4 Métricas de avaliação

As métricas de avaliação utilizadas foram o Erro Médio Absoluto, do inglês *Mean Absolute Error* (MAE), dado pela Equação 4.3, e o Erro Médio Quadrático, do inglês *Mean Squared Error* (MSE), dado pela Equação 4.4. O MAE representa a soma das diferenças absolutas e MSE a soma do quadrado das diferenças.

$$MAE = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} |C_i - C_i^T|^2 \tag{4.3}$$

$$MSE = \sqrt{\frac{1}{N_{te}} \sum_{i=1}^{N_{te}} |C_i - C_i^T|^2} \tag{4.4}$$

Nestas equações, N_{te} corresponde ao número de imagens de testes, C_i^T é a número total de anotações verdadeiras e C_i é o número total estimado (por meio da integração do mapa produzido pela rede) pela Equação 4.5, na qual L e W são, respectivamente, comprimento e largura do mapa de densidade e *z(l,w)* corresponde ao valor do pixel na posição (l, w).

$$C_i = \sum_{l=1}^{L} \sum_{w=1}^{W} z(l, w) \tag{4.5}$$

4.2 Etapa de Testes

Nesta etapa, cada subconjunto de amostras de testes será aplicado a cada um dos modelos propostos neste trabalho. A ideia é avaliar o comportamento e a capacidade de generalização de cada um deles de acordo com os resultados alcançados caracterizados pelas métricas previamente definidas.

O modelo deverá ser capaz de, a partir de uma imagem de multidão, produzir o seu respectivo mapa de densidade e com ele permitir a estimativa do número de pessoas na cena, ou seja, a contagem da multidão.

4.3 Resultados Alcançados

Nas próximas subseções serão detalhados e analisados os resultados alcançados por cada uma das abordagens propostas.

4.3.1 Ca-CCNet

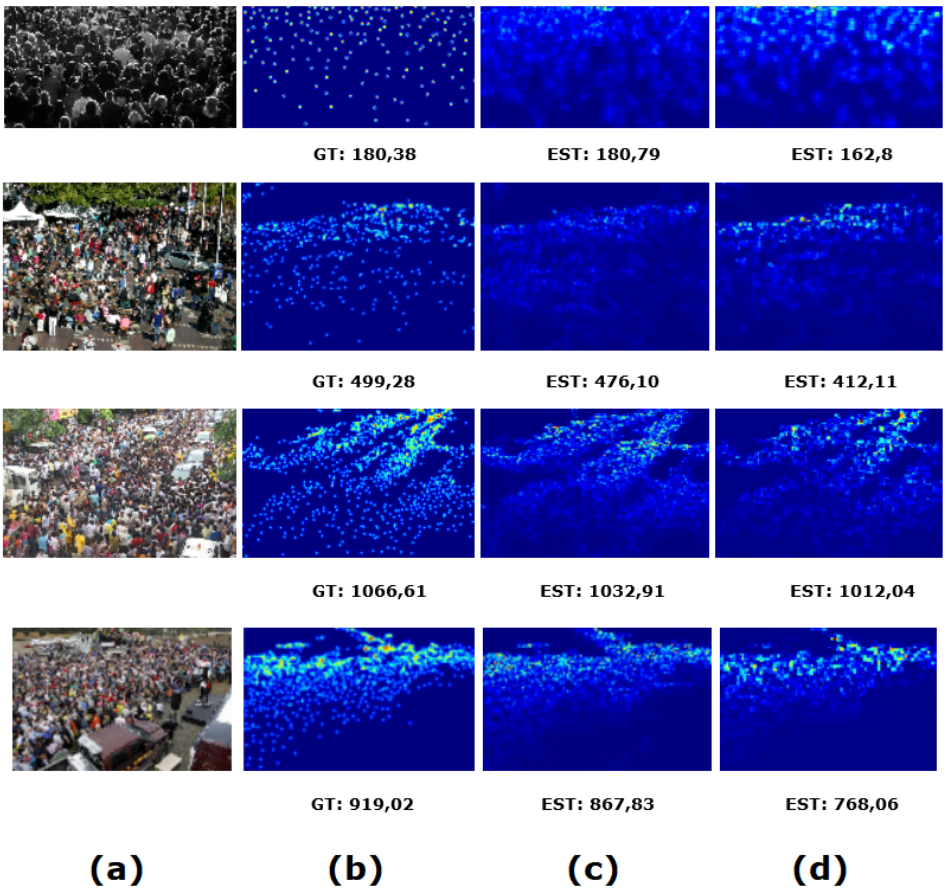
Esta proposta consiste em uma nova arquitetura baseada em CapsNet para lidar com a tarefa de contagem de multidões e estimativa de mapas de densidade. Experimentos demonstraram que a redução da resolução das imagens e também do número de iterações do algoritmo de roteamento dinâmico, melhoram sensivelmente o desempenho na etapa de treinamento. Também foi demonstrado que o modelo baseado no CapsNet proposto supera algumas imagens de teste de referência, além de utilizar um número reduzido de parâmetros, quando comparado a outras abordagens, como por exemplo a CSRNet (Li; Zhang; Chen, 2018). Nas Figuras 27 e 28 temos os resultados qualitativos desta proposta com a base ShanghaiTech Parte A e Parte B, respectivamente.

Os resultados quantitativos estão contidos nas Tabelas 6 e 7. Para as imagens de treinamento do repositório ShanghaiTech, os resultados são melhores do que as abordagens propostas por Zhang et al. (2015) e Zhang et al. (2016). Na Tabela 8, estão os resultados alcançados na base UCF_CC_50.

Esta é uma rede promissora, porque os mapas de densidade produzidos são semelhantes ao *ground-truth*. No entanto, o resultado não foi o esperado e o desempenho não é satisfatório principalmente em imagens em escala de cinza ou imagens de multidões muito densas quando comparadas aos métodos mais recentes baseados em CNN tradicionais.

Em algumas imagens em escala de cinza, a textura de certas regiões, como árvores e vegetação, foram classificadas como regiões de multidões, causando falsos positivos, de acordo com a Figura 29. Esperava-se um melhor desempenho ao lidar com este desafio.

Figura 27 – Alguns resultados alcançados com a proposta Ca-CCNet nas amostras de testes da base ShanghaiTech Parte A. Imagem original da multidão (a); *ground-truth* (GT) (b); mapa de densidade estimado pela nossa proposta (c); e comparação com outra proposta CSRNet (d).



Fonte: Elaborada pelo autor

Tabela 6 – Resultados alcançados com a proposta Ca-CCNet em comparação com a literatura no conjunto de teste ShanghaiTech Parte A.

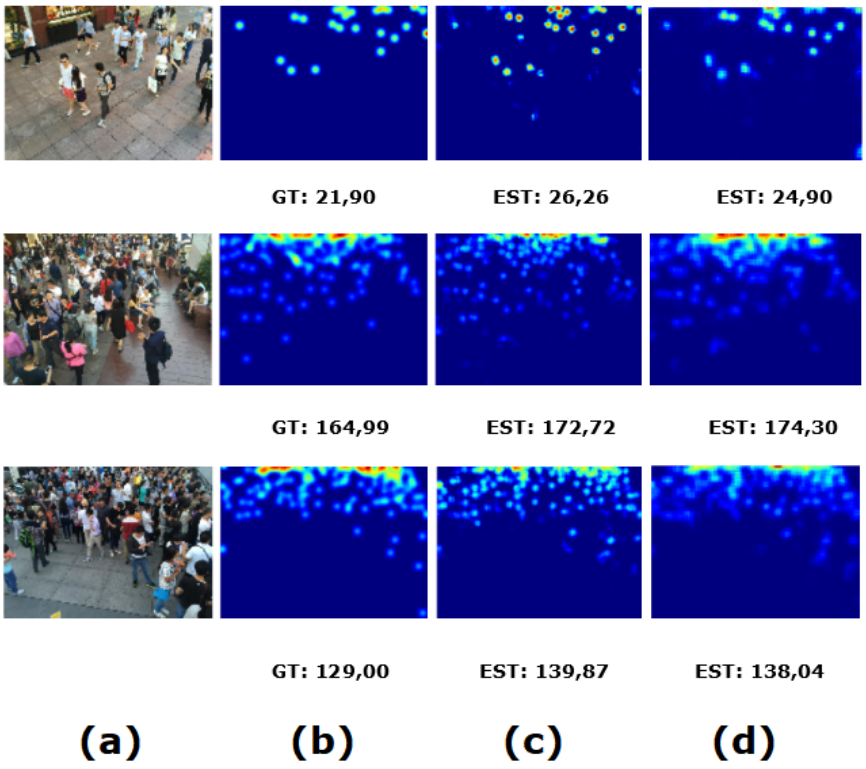
Métodos	MAE	MSE
Zhang et al. (2015)	181,8	277,7
Zhang et al. (2016)	110,2	173,2
Ca-CCNet	**106,0**	**167,6**
Sindagi e Patel (2017)	101,3	152,4
Li, Zhang e Chen (2018)	68,2	115,0
Shi et al. (2018)	66,3	106,4
Sindagi e Patel (2020)	62,9	94,9
Shi et al. (2018) + Li, Zhang e Chen (2018)	62,4	102,0

Fonte: Elaborada pelo autor

4.3.2 QU-CCNet

Nesta proposta foi desenvolvida uma CNN inspirada em arquitetura em "U" que oferece um tipo de previsão densa e tenta capturar o contexto através da combinação de informações extraídas de diferentes camadas da rede.

Figura 28 – Alguns resultados alcançados com a proposta Ca-CCNet nas amostras de testes da base ShanghaiTech Parte B. Imagem original da multidão (a); *ground-truth* (GT) (b); mapa de densidade estimado pela nossa proposta (c); e comparação com outra proposta CSRNet (d).



Fonte: Elaborada pelo autor

Tabela 7 – Resultados alcançados com a proposta Ca-CCNet em comparação com a literatura no conjunto de teste ShanghaiTech Parte B.

Métodos	MAE	MSE
Zhang et al. (2015)	32,0	49,8
Zhang et al. (2016)	26,4	41,3
Ca-CCNet	**21,1**	**39,2**
Sindagi e Patel (2017)	20,0	31,1
Li, Zhang e Chen (2018)	10,6	16,0
Shi et al. (2018)	8,9	13,5
Sindagi e Patel (2020)	8,1	13,4
Shi et al. (2018) + Li, Zhang e Chen (2018)	7,6	11,8

Fonte: Elaborada pelo autor

Esta proposta tem parte da rede inicializada pelos pesos da VGG-16. Com isso, percebemos que ela rapidamente (com poucas iterações no treinamento) alcança a capacidade de produzir mapas de densidade, tornando a etapa de treinamento mais curta do ponto de vista do número de iterações.

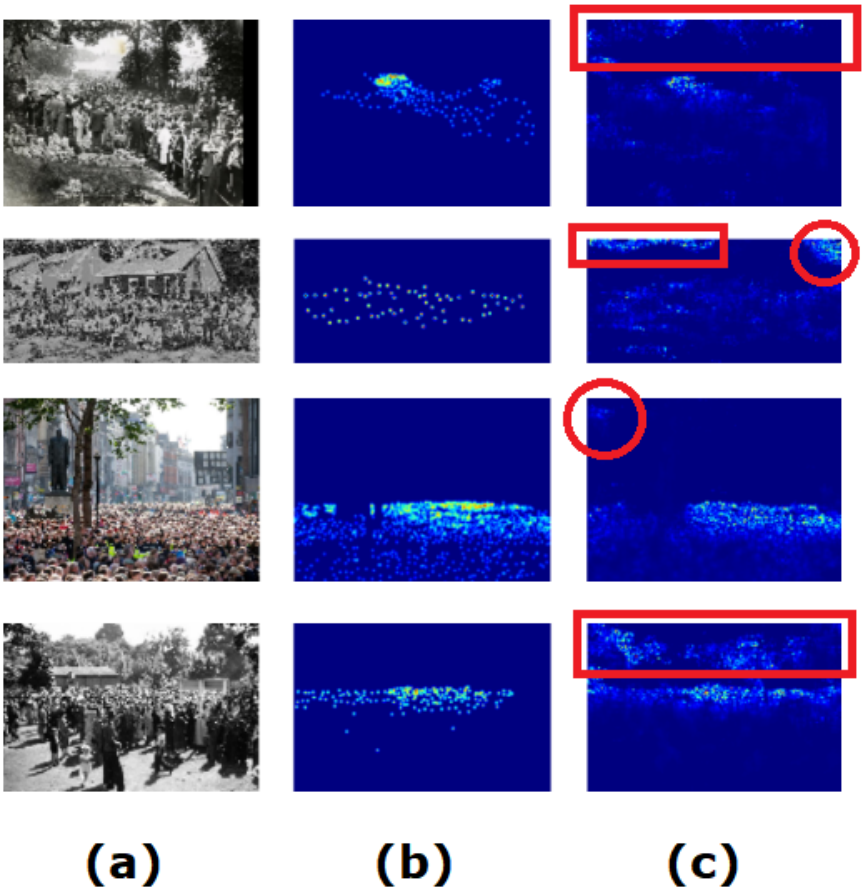
Na Figura 30 estão os mapas produzidos por esta proposta nas amostras de testes da base ShanghaiTech Parte A e na Figura 31 as amostras de testes da Parte B do mesmo repositório.

Tabela 8 – Resultados alcançados com a proposta Ca-CCNet em comparação com o literatura na base UCF_CC_50.

Métodos	MAE	MSE
Idrees et al. (2013)	419,5	541,6
Zhang et al. (2015)	467,0	498,5
Zhang et al. (2016)	377,6	509,1
Ca-CCNet	**372,4**	**532,9**
Sindagi e Patel (2017)	322,8	397,9
Shi et al. (2018)	267,9	357,8
Li, Zhang e Chen (2018)	266,1	397,5
Sindagi e Patel (2020)	256,2	348,4
Shi et al. (2018) + Li, Zhang e Chen (2018)	241,7	320,7

Fonte: Elaborada pelo autor

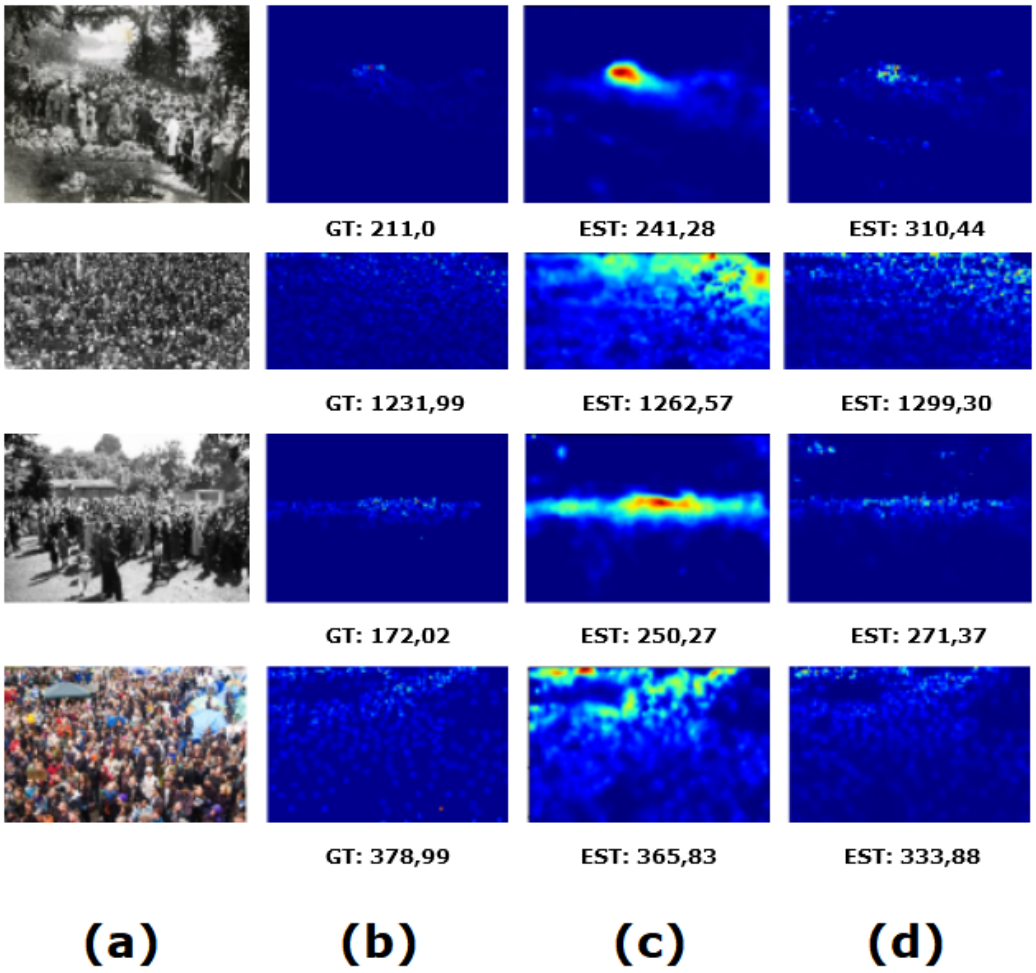
Figura 29 – Alguns fracos resultados alcançados pela proposta Ca-CCNet com as amostras de testes da base ShanghaiTech Parte A. A coluna (a) contém imagem original da multidão, (b) contém o *ground truth* e (c) contém o mapa de densidade estimado pela nossa proposta com falsos positivos destacados em vermelho.



Fonte: Elaborada pelo autor

Podemos considerar esta rede um pouco mais pesada do ponto de vista do número de camadas em comparação com as outras propostas, e consequentemente maior também em relação ao parâmetros ($53M$).

Figura 30 – Alguns resultados alcançados com a proposta QU-CCNet nas amostras de testes da base ShanghaiTech Parte A. Imagem original da multidão (a); *ground-truth* (GT) (b); mapa de densidade estimado pela nossa proposta (c); e comparação com outra proposta CSRNet (d).



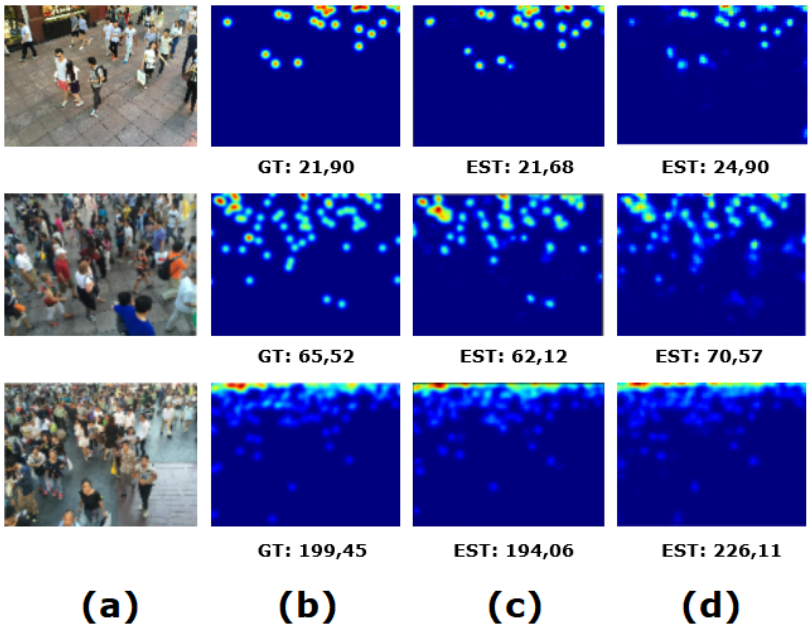
Fonte: Elaborada pelo autor

Os resultados quantitativos alcançados com as amostras de testes da base ShanghaiTech, partes A e B, estão descritos nas Tabelas 9 e 10, respectivamente. Já os resultados alcançados com a base UCF_CC_50 estão na Tabela 11.

Os resultados quantitativos não foram tão ruins, superando algumas abordagens. Esperava-se que a utilização de um decodificador pudesse minimizar o problema da perda da localização das *features*. Porém, esperava-se um resultado quantitativo melhor devido ao sucesso alcançado na tarefa original (U-Net), mesmo que em outro tipo de tarefa.

Os mapas produzidos por esta abordagem costumam ser, visivelmente, mais densos se comparados com o *ground-truth*. Isto sugere uma estimativa de contagem maior do que o número oficial de indivíduos. No entanto, apesar dos mapas serem mais densos, o número total de indivíduos estimados tende a diminuir quando estendida a etapa de treinamento.

Figura 31 – Alguns resultados alcançados com a proposta QU-CCNet nas amostras de testes da base ShanghaiTech Parte B. Imagem original da multidão (a); *ground-truth* (GT) (b); mapa de densidade estimado pela nossa proposta (c); e comparação com outra proposta CSRNet (d).



Fonte: Elaborada pelo autor

Tabela 9 – Resultados alcançados com a proposta QU-CCNet em comparação com a literatura no conjunto de teste ShanghaiTech Parte A.

Métodos	MAE	MSE
Zhang et al. (2015)	181,8	277,7
Zhang et al. (2016)	110,2	173,2
Weng e Lin (2018)	108,2	171,3
Sindagi e Patel (2017)	101,3	152,4
QU-CCNet	**75,5**	**119,4**
Li, Zhang e Chen (2018)	68,2	115,0
Shi et al. (2018)	66,3	106,4
Sindagi e Patel (2020)	62,9	94,9
Shi et al. (2018) + Li, Zhang e Chen (2018)	62,4	102,0

Fonte: Elaborada pelo autor

Exceto por algumas poucas imagens, principalmente em escala de cinza, este modelo tem melhor capacidade de diferenciar texturas que correspondem a árvores e vegetação de texturas que correspondem a multidões, reduzindo os falsos positivos.

4.3.3 CaTL-CCNet

Os resultados alcançados pela abordagem híbrida, que combina parte de uma CNN previamente treinada explorando a transferência de aprendizado com parte da rede em cápsula, na base ShanghaiTech, estão descritos nas Tabelas 12 e 13.

Tabela 10 – Resultados alcançados com a proposta QU-CCNet em comparação com a literatura no conjunto de teste ShanghaiTech Parte B.

Métodos	MAE	MSE
Zhang et al. (2015)	32,0	49,8
Zhang et al. (2016)	26,4	41,3
Sindagi e Patel (2017)	20,0	31,1
Li, Zhang e Chen (2018)	10,6	16,0
QU-CCNet	**10,2**	**15,9**
Shi et al. (2018)	8,9	13,5
Sindagi e Patel (2020)	8,1	13,4
Shi et al. (2018) + Li, Zhang e Chen (2018)	7,6	11,8

Fonte: Elaborada pelo autor

Tabela 11 – Resultados alcançados com a proposta QU-CCNet em comparação com a literatura na base UCF_CC_50.

Métodos	MAE	MSE
Idrees et al. (2013)	419,5	541,6
Zhang et al. (2015)	467,0	498,5
Zhang et al. (2016)	377,6	509,1
QU-CCNet	**333,6**	**438,0**
Sindagi e Patel (2017)	322,8	397,9
Shi et al. (2018)	267,9	357,8
Li, Zhang e Chen (2018)	266,1	397,5
Sindagi e Patel (2020)	256,2	348,4
Shi et al. (2018) + Li, Zhang e Chen (2018)	241,7	320,7

Fonte: Elaborada pelo autor

Quando se trata do resultado, esta proposta superou alguns dos importantes trabalhos disponíveis na literatura para as imagens de testes do repositório ShanghaiTech Parte B e ficou muito próximo de superar na Parte A. Os resultados com a base UCF_CC_50 estão na Tabela 14.

Esta proposta visa combinar nas camadas de cápsulas as características extraídas das camadas convolucionais anteriores, na tentativa de melhor distinguir quais estímulos pertencem ao objeto de análise, no nosso caso a multidão, e quais pertencem a outros como o plano de fundo. Esta estratégia se mostrou eficiente, levando em consideração os resultados alcançados. Os resultados qualitativos, ou seja, os mapas produzidos, estão representados na Figuras 32 e 33 (ShanghaiTech Parte A e ShanghaiTech Parte B, respectivamente).

Apesar da questão conceitual relacionada a crítica das camadas de *pooling*, a combinação de algumas poucas destas camadas com cápsulas na prática, para esta tarefa de estimativa de contagem de multidões, se mostrou aceitável e os resultados alcançados foram satisfatórios e superaram importantes abordagens.

Quando analisamos os mapas de forma qualitativa, percebemos que esta proposta produz mapas mais precisos e coerentes com o *ground-truth* em comparação com outros trabalhos, por exemplo a CSRNet (Li; Zhang; Chen, 2018). Provavelmente, devido ao

uso de convoluções dilatadas, determinadas regiões ficam mais "borradas" do que o mapa original, causando imprecisão em alguns mapas, conforme ilustra a Figura 34.

4.4 Comparação das abordagens

Na Tabela 15 foram apresentados os resultados alcançados por todas as três propostas em comparação com a literatura na base ShanghaiTech. Já os resultados alcançados com a base UCF_CC_50 estão na Tabela 16.

Podemos perceber que a proposta CaTL-CCNet alcançou os melhores resultados, indicando que esta arquitetura que explora a transferência de aprendizado de uma CNN, combinada com camadas em cápsulas no final da rede, seria a maneira mais adequada para este tipo de tarefa.

Uma entidade no campo visual periférico é muito mais difícil de identificar na presença de outras entidades muito próximas, fenômeno conhecido como "aglomeração" (PELLI; PALOMARES; MAJAJ, 2005) e a proximidade pode tornar a identificação ambígua e confusa com os vizinhos mais próximos. Talvez este fenômeno justifique os resultados não tão bons em bases com imagens de multidões muito densas, onde os indivíduos estão muito próximos e difíceis de serem identificados isoladamente.

O fenômeno da "aglomeração" também ocorre entre as partes de um objeto. Essa aglomeração interna também prejudica gravemente a percepção e a identificação.

Martelli, Majaj e Pelli (2005) mostram que os observadores processam palavras e rostos da mesma maneira e os efeitos de familiaridade e "aglomeração" não distinguem entre si, portanto palavras e rostos são reconhecidos por partes, e suas partes (letras e feições faciais) são reconhecidas holisticamente.

Esperava-se um desempenho melhor das cápsulas ao lidar com a relação semântica e hierárquica entre as partes identificadas de uma entidade, neste caso os indivíduos da

Tabela 12 – Resultados alcançados com a proposta CaTL-CCNet em comparação com a literatura no conjunto de teste ShanghaiTech Parte A.

Métodos	MAE	MSE
Zhang et al. (2015)	181,8	277,7
Zhang et al. (2016)	110,2	173,2
Weng e Lin (2018)	108,2	171,3
Sindagi e Patel (2017)	101,3	152,4
CaTL-CCNet	**69,5**	**114,4**
Li, Zhang e Chen (2018)	68,2	115,0
Shi et al. (2018)	66,3	106,4
Sindagi e Patel (2020)	62,9	94,9
Shi et al. (2018) + Li, Zhang e Chen (2018)	62,4	102,0

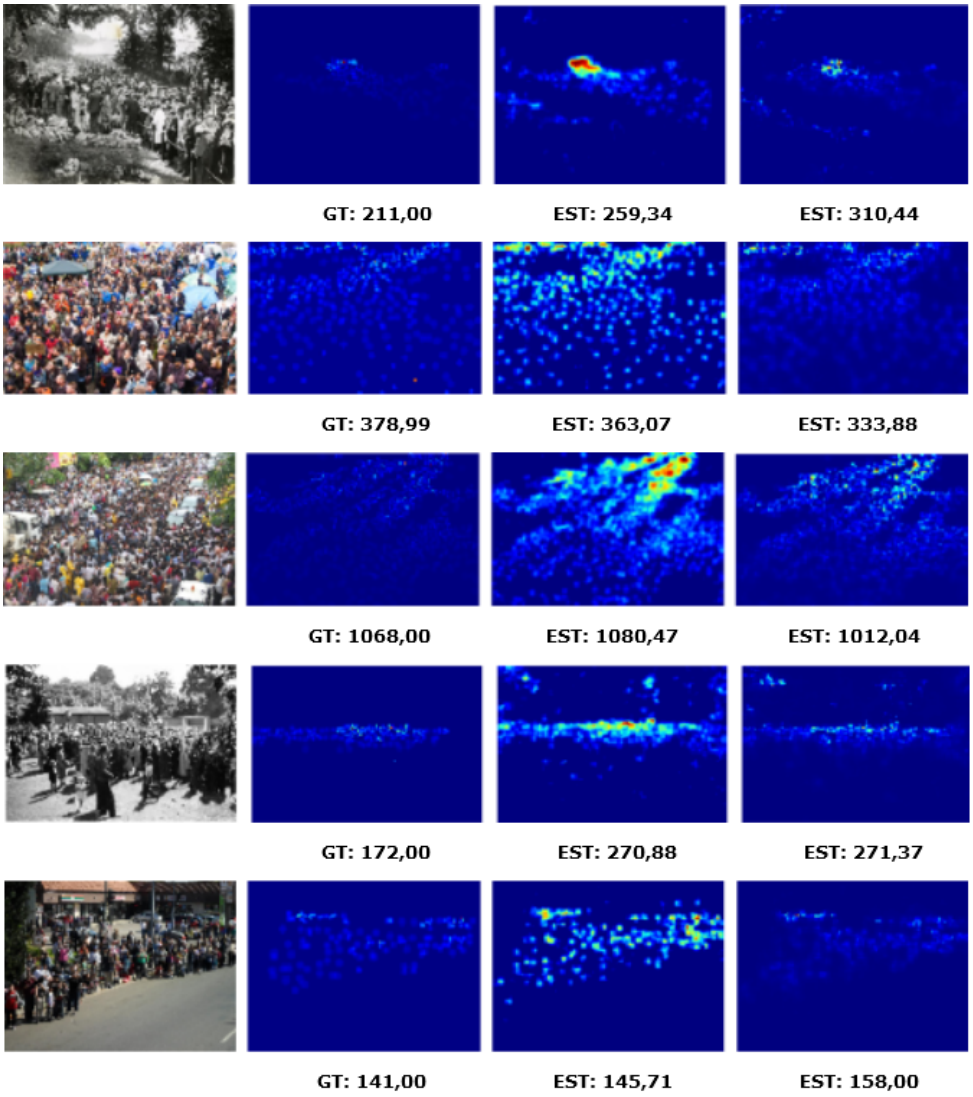
Fonte: Elaborada pelo autor

multidão. No entanto, apesar dos obstáculos nas cenas de multidões densas, nas imagens de multidões esparsas, onde os indivíduos não estão muito próximos uns dos outros, os resultados foram melhores.

Nas Figuras 35 e 36 estão os resultados qualitativos representados pelos mapas de densidade de algumas amostras de testes da base ShanghaiTech (Parte A e Parte B, respectivamente).

O comportamento das propostas foi semelhante também na base UCF_CC_50, ou seja, os melhores resultados foram alcançados com a proposta CaTL-CCNet.

Figura 32 – Alguns resultados alcançados com a proposta CaTL-CCNet nas amostras de testes da base ShanghaiTech Parte A. Imagem original da multidão (a); *ground-truth* (GT) (b); mapa de densidade estimado pela nossa proposta (c); e comparação com outra proposta CSRNet (d).



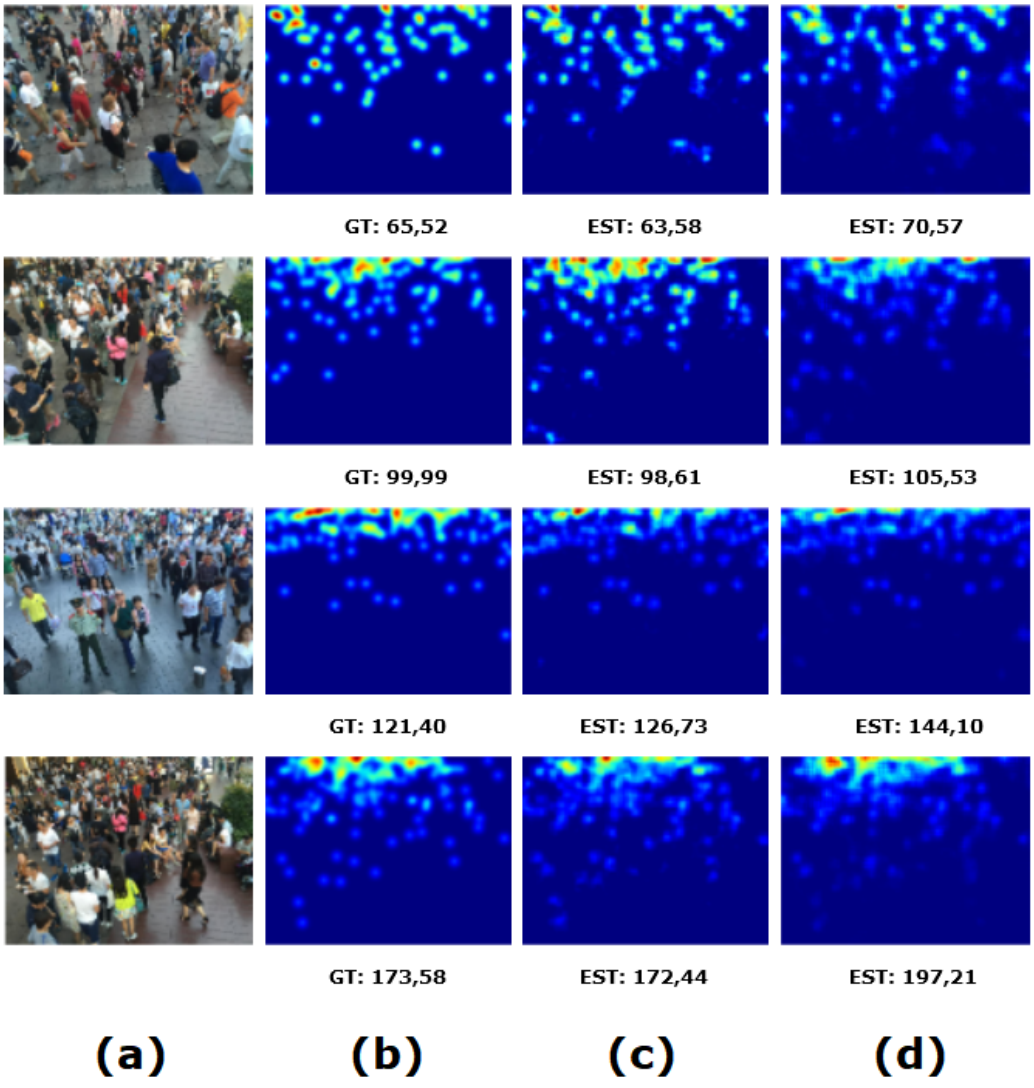
Fonte: Elaborada pelo autor

Tabela 13 – Resultados alcançados com a proposta CaTL-CCNet em comparação com a literatura no conjunto de teste ShanghaiTech Parte B.

Métodos	MAE	MSE
Zhang et al. (2015)	32,0	49,8
Zhang et al. (2016)	26,4	41,3
Sindagi e Patel (2017)	20,0	31,1
Li, Zhang e Chen (2018)	10,6	16,0
Shi et al. (2018)	8,9	13,5
CaTL-CCNet	**8,6**	**16,7**
Sindagi e Patel (2020)	8,1	13,4
Shi et al. (2018) + Li, Zhang e Chen (2018)	7,6	11,8

Fonte: Elaborada pelo autor

Figura 33 – Alguns resultados alcançados com a proposta CaTL-CCNet nas amostras de testes da base ShanghaiTech Parte B. Imagem original da multidão (a); *ground-truth* (GT) (b); mapa de densidade estimado pela nossa proposta (c); e comparação com outra proposta CSRNet (d).



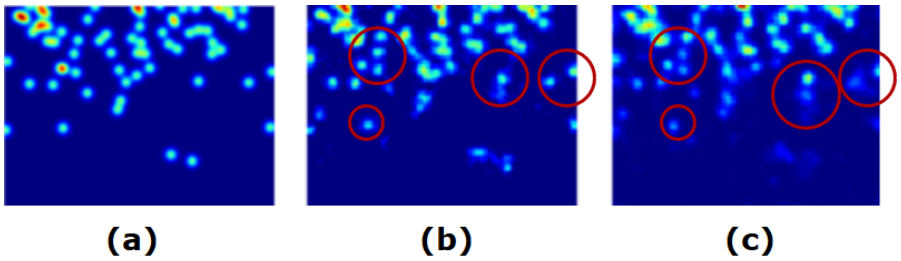
Fonte: Elaborada pelo autor

Tabela 14 – Resultados alcançados com a proposta CaTL-CCNet em comparação com a literatura na base UCF_CC_50.

Métodos	MAE	MSE
Idrees et al. (2013)	419,5	541,6
Zhang et al. (2015)	467,0	498,5
Zhang et al. (2016)	377,6	509,1
Sindagi e Patel (2017)	322,8	397,9
CaTL-CCNet	**303,1**	**407,5**
Shi et al. (2018)	267,9	357,8
Li, Zhang e Chen (2018)	266,1	397,5
Sindagi e Patel (2020)	256,2	348,4
Shi et al. (2018) + Li, Zhang e Chen (2018)	241,7	320,7

Fonte: Elaborada pelo autor

Figura 34 – Comparação entre o mapa original (a); proposta CaTL-CCNet (b); e a abordagem CSRNet (c). Percebemos que o uso de convoluções dilatadas torna algumas regiões do mapa mais "borradas".



Fonte: Elaborada pelo autor

No entanto, vale ressaltar que no caso da base UCF_CC_50 o treinamento foi realizado com validação cruzada. Neste caso, nas propostas QU-CCNet e CaTL-CCNet, percebemos que os resultados foram prejudicados devido a um dos subconjuntos que alcançou resultados muito ruins, formado por imagens de multidões muito densas, jogando a média da estimativa para um valor muito alto.

Apesar da diferença na contagem e do erro relativamente alto como um todo na base UCF_CC_50, alguns mapas ficaram semelhantes ao *ground-truth*, tornando os resultados em algumas imagens eficiente tanto do ponto de vista quantitativo quanto qualitativo, conforme ilustra a Figura 37.

Tabela 15 – Resultados alcançados pelas três propostas em comparação com a literatura.

Métodos	Parte A		Parte B	
	MAE	MSE	MAE	MSE
Zhang et al. (2015)	181,8	277,7	32,0	49,8
Zhang et al. (2016)	110,2	173,2	26,4	41,3
Sindagi e Patel (2017)	101,3	152,4	20,0	31,1
Sam, Surya e Babu (2017)	90,4	135,0	-	-
Zeng et al. (2017)	83,8	127,4	-	-
Li, Zhang e Chen (2018)	68,2	115,0	10,6	16,0
Shi et al. (2018)	66,3	106,4	8,9	13,5
Sindagi e Patel (2020)	62,9	94,9	8,1	13,4
Shi et al. (2018) + Li, Zhang e Chen (2018)	62,4	102,0	7,6	11,8
Ca-CCNet	**106,0**	**167,6**	**24,0**	**37,4**
QU-CCNet	**75,5**	**119,4**	**10,2**	**15,9**
CaTL-CCNet	**69,5**	**119,4**	**8,6**	**16,7**

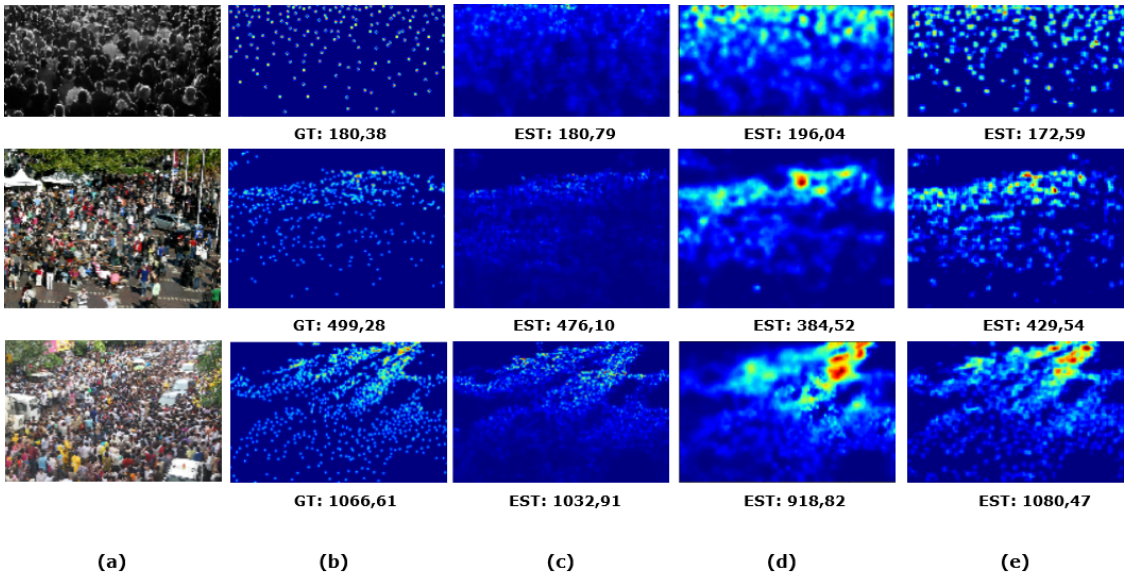
Fonte: Elaborada pelo autor

Tabela 16 – Resultados alcançados com a proposta CaTL-CCNet em comparação com o literatura na base UCF_CC_50.

Métodos	MAE	MSE
Idrees et al. (2013)	419,5	541,6
Zhang et al. (2015)	467,0	498,5
Zhang et al. (2016)	377,6	509,1
Sindagi e Patel (2017)	322,8	397,9
Shi et al. (2018)	267,9	357,8
Li, Zhang e Chen (2018)	266,1	397,5
Sindagi e Patel (2020)	256,2	348,4
Shi et al. (2018) + Li, Zhang e Chen (2018)	241,7	320,7
Ca-CCNet	**372,4**	**532,9**
QU-CCNet	**333,6**	**438,0**
CaTL-CCNet	**303,1**	**407,5**

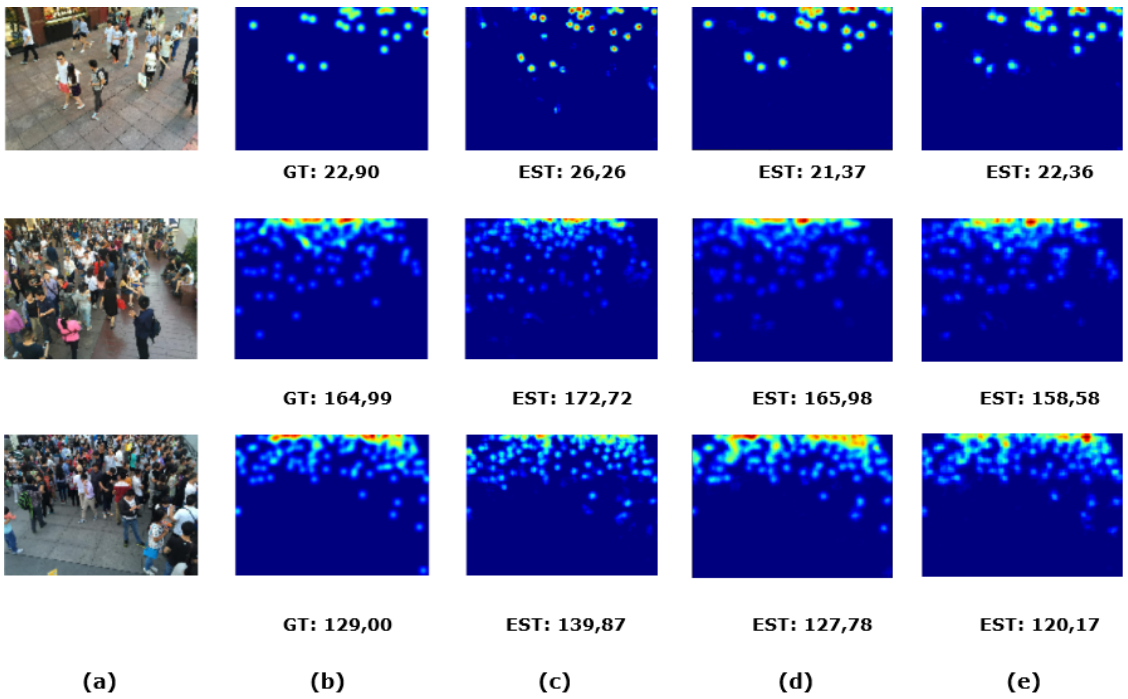
Fonte: Elaborada pelo autor

Figura 35 – Comparação entre as propostas com amostras de testes Shanghai-Tech Parte A imagem original (a); *ground-truth* (GT) (b); proposta Ca-CCNet (c); proposta QU-CCNet (d); e proposta CaTL-CCNet (e).



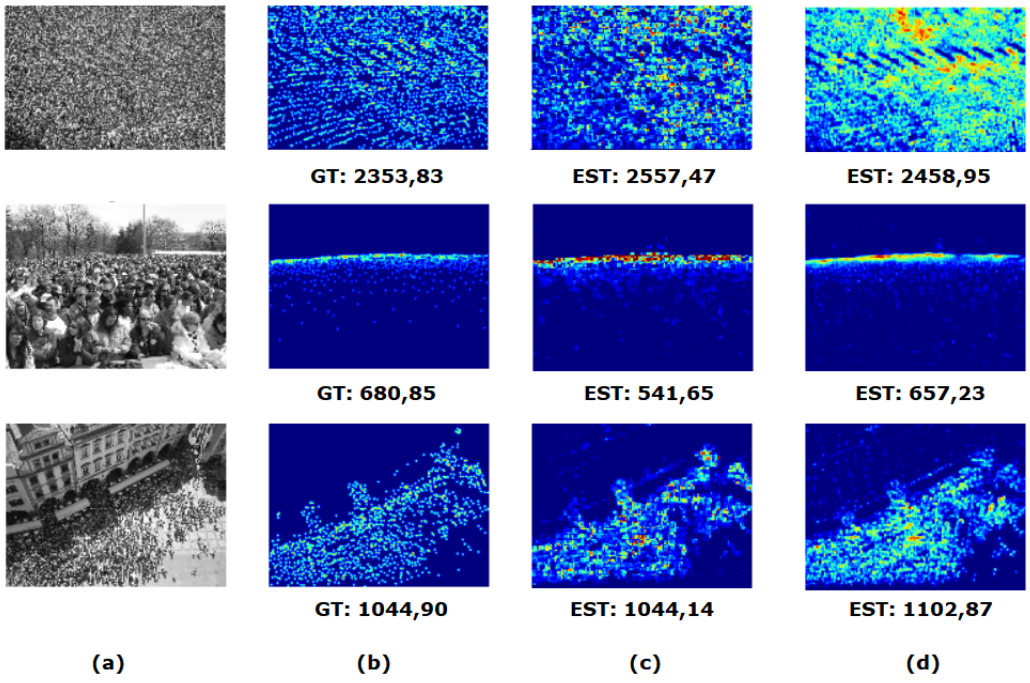
Fonte: Elaborada pelo autor

Figura 36 – Comparação entre as propostas com amostras de testes Shanghai-Tech Parte B imagem original (a); *ground-truth* (GT) (b); proposta Ca-CCNet (c); proposta QU-CCNet (d); e proposta CaTL-CCNet (e).



Fonte: Elaborada pelo autor

Figura 37 – Comparação entre as propostas com amostras da base UCF_CC_50. Imagem original (a); *ground-truth* (GT) (b); proposta CaTL-CCNet (c); e proposta QU-CCNet (d).



Fonte: Elaborada pelo autor

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propôs a implementação e análise de redes neurais artificiais profundas para a solução do problema da contagem de multidões, que é uma das tarefas da área de visão computacional.

O problema da contagem de multidões pode ser solucionado computacionalmente de diferentes formas, no entanto, os melhores resultados têm sido alcançados usando métodos baseados em estimativa de densidade. Por este motivo, as três propostas deste trabalho foram baseadas em mapas de densidade.

Uma das propostas chamada de QU-CCNet, consistiu de uma CNN que teve sua arquitetura desenvolvida de forma semelhante ao formado em "U", baseado no fato de que arquiteturas neste formato alcançaram bons resultados em tarefas de segmentação de imagens e esperávamos o mesmo sucesso na solução do problema da estimativa da contagem de multidões.

Percebemos que com poucas iterações na etapa de treinamento a rede já é capaz de produzir os mapas de densidade. No entanto, esta proposta necessita de uma estratégia diferenciada de aumento dos dados, pois ela tem uma forte tendência ao *overfitting*.

Os resultados alcançados com a proposta QU-CCNet se mostraram regulares e tornaram a rede promissora quando comparamos com a literatura, pois esta proposta superou alguns trabalhos que utilizaram abordagens semelhantes. Para trabalhos futuros, pretendemos realizar ajustes nos hiperparâmetros e novas estratégias de aumento dos dados na expectativa de melhorar os resultados.

Alguns problemas encontrados nas CNNs também motivaram novas abordagens, como é o caso das CapsNets. Com este tipo de rede, inspiradas na ideia de cápsulas para melhor modelar as relações hierárquicas e solucionar o problema da perda de informações da localização das características ao longo da rede, esperávamos alcançar o estado da arte na contagem de multidões. No entanto, isto não foi totalmente possível.

Na primeira proposta chamada de Ca-CCNet os resultados não estão de acordo com a expectativa gerada sobre esta arquitetura de rede CapsNet. Esta abordagem não se mostrou satisfatória na medida em que os resultados quantitativos alcançados não foram os esperados, apesar dos resultados qualitativos estarem, relativamente, coerentes com outros resultados encontrados na literatura.

Suspeita-se ainda de que seja necessária uma revisão de alguns dos hiperparâmetros

utilizados na rede, como por exemplo o tamanho do *kernel* e o uso de camadas convolucionais dilatadas que pode ter sido importante em alguns trabalhos, mas talvez não nesta proposta. Além disso, o algoritmo de roteamento dinâmico é mais caro e lento e pode interferir também no desempenho do algoritmo *backpropagation*.

No final das contas, consideramos uma rede promissora na medida que consegue construir mapas de densidade semelhantes ao *ground-truth*, chegando a ser mais eficiente em alguns casos. Com os ajustes necessários e maior investimento em treinamento, talvez possa se tornar mais eficiente do ponto de vista dos resultados.

Por fim, a terceira e última proposta chamada CaTL-CCNet superou importantes trabalhos em contagem de multidões na base ShanguaiTech Parte B e ficou próxima do feito também na Parte A.

O início da rede possui camadas e pesos extraídos de uma rede VGG-19, que no final é combinada com camadas em cápsulas. Este é um *framework* amplamente utilizado neste tipo de tarefa.

A proposta desta rede é de usar as cápsulas para melhor entender e representar *features* extraídas das camadas anteriores e tomar melhores decisões no final da rede para produzir mapas de densidade com qualidade, ou seja, mais precisos e mais próximos do *ground-truth*.

Se compararmos as métricas de avaliação, a proposta CaTL-CCNet alcançou os melhores resultados. Além disso, é uma rede fácil de treinar e que devido a transferência de aprendizado tem o tempo de treinamento baixo e não foram necessárias muitas iterações nesta etapa. Consideramos esta como a melhor e mais promissora proposta do ponto de vista de implementação e dos resultados alcançados.

Para trabalhos futuros pretendemos seguir algumas direções. A primeira diz respeito ao estudo e aplicação de redes em cápsulas em outras tarefas como segmentação e classificação de imagens digitais. Nesta mesma direção, identificar e propor melhorias no algoritmo de roteamento dinâmico. Uma segunda direção vai ao encontro de novas soluções para o problema da estimativa de contagem com mapas de densidade utilizando redes previamente treinadas, uma abordagem que vem obtendo sucesso, combinadas com novas estratégias como, por exemplo, morfologia matemática.

REFERÊNCIAS

ABADI, M. et al. Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. USA: USENIX Association, 2016. (OSDI'16), p. 265–283. ISBN 9781931971331.

BOOMINATHAN, L.; KRUTHIVENTI, S. S. S.; BABU, R. V. Crowdnet: A deep convolutional network for dense crowd counting. In: Proceedings of the 24th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2016. (MM '16), p. 640–644. ISBN 9781450336031. Disponível em: <https://doi.org/10.1145/2964284.2967300>.

CIREGAN, D.; MEIER, U.; SCHMIDHUBER, J. Multi-column deep neural networks for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2012. p. 3642–3649.

FUKUSHIMA, K.; MIYAKE, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: AMARI, S.-i.; ARBIB, M. A. (Ed.). Competition and Cooperation in Neural Nets. Berlin, Heidelberg: Springer Berlin Heidelberg, 1982. p. 267–285. ISBN 978-3-642-46466-9.

HAYKIN, S. S. Neural networks and learning machines. Third. Upper Saddle River, NJ: Pearson Education, 2009.

He, K. et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV). [S.l.: s.n.], 2015. p. 1026–1034.

He, K. et al. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2016. p. 770–778.

HINTON, G. E.; SABOUR, S.; FROSST, N. Matrix capsules with EM routing. In: International Conference on Learning Representations. [s.n.], 2018. Disponível em: <https://openreview.net/forum?id=HJWLfGWRb>.

HUBEL, D. H.; WIESEL, T. N. Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology, v. 148, p. 574–91, 1959.

IDREES, H. et al. Multi-source multi-scale counting in extremely dense crowd images. In: 2013 IEEE CVPR. [S.l.: s.n.], 2013. p. 2547–2554. ISSN 1063-6919.

JACOBS, H. To count a crowd. Columbia Journalism Review, Columbia University, Graduate School of Journalism, v. 6, n. 1, p. 37, 1967.

JUNIOR, J. C. S. J.; MUSSE, S. R.; JUNG, C. R. Crowd analysis using computer vision techniques. IEEE Signal Processing Magazine, IEEE, v. 27, n. 5, p. 66–77, 2010.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. In: BENGIO, Y.; LECUN, Y. (Ed.). 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. [s.n.], 2015. Disponível em: <http://arxiv.org/abs/1412.6980>.

KOWCIKA, A.; SRIDHAR, S. A literature study on crowd (people) counting with the help of surveillance videos. International Journal of Innovative Technology and Research, p. 2353–2361, 2016.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: NIPS. [S.l.: s.n.], 2012.

KUNCHEVA, L. Combining Pattern Classifiers: Methods and Algorithms. 2nd. ed. [S.l.]: Wiley Publishing, 2014. ISBN 1118315235.

LALONDE, R.; BAGCI, U. Capsules for object segmentation. CoRR, abs/1804.04241, 2018. Disponível em: <http://arxiv.org/abs/1804.04241>.

LECUN, Y. et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, v. 86, p. 2278 – 2324, 12 1998.

LEMPITSKY, V.; ZISSERMAN, A. Learning to count objects in images. In: Advances in Neural Information Processing Systems 23. [S.l.: s.n.], 2010. p. 1324–1332.

LI, M. et al. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: 2008 19th ICPR. [S.l.: s.n.], 2008. p. 1–4. ISSN 1051-4651.

Li, Y.; Zhang, X.; Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2018. p. 1091–1100.

LIU, L. et al. Crowd counting using deep recurrent spatial-aware network. In: LANG, J. (Ed.). Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. ijcai.org, 2018. p. 849–855. Disponível em: <https://doi.org/10.24963/ijcai.2018/118>.

LOY, C. C. et al. Crowd counting and profiling: Methodology and evaluation. In: ALI, S. et al. (Ed.). Modeling, Simulation and Visual Analysis of Crowds - A Multidisciplinary Perspective. Springer, 2013, (The International Series in Video Computing, v. 11). p. 347–382. Disponível em: <https://doi.org/10.1007/978-1-4614-8483-7_14>.

Luo, Y.; Lu, J.; Zhang, B. Crowd counting for static images: A survey of methodology. In: 2020 39th Chinese Control Conference (CCC). [S.l.: s.n.], 2020. p. 6602–6607.

MARTELLI, M.; MAJAJ, N.; PELLI, D. Are faces processed like words? a diagnostic test for recognition by parts. Journal of vision, v. 5, p. 58–70, 02 2005.

OH, M.; OLSEN, P. A.; RAMAMURTHY, K. N. Crowd counting with decomposed uncertainty. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 2020. p. 11799–11806. Disponível em: <https://aaai.org/ojs/index.php/AAAI/article/view/6852>.

PEDRINI, H.; SCHWARTZ, W. R. Análise de Imagens Digitais: Princípios, Algoritmos e Aplicações. [S.l.]: Editora Thomson Learning, 2007. 528 p. ISBN 978-85-221-0595-3.

PELLI, D.; PALOMARES, M.; MAJAJ, N. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. Journal of vision, v. 4, p. 1136–69, 01 2005.

PERRETT, D.; ORAM, M. Neurophysiology of shape processing. Image and Vison Computing, v. 11, n. 6, p. 317–333, jul. 1993.

PRATHIBA, G. T.; DHAS, Y. Literature survey for people counting and human detection. IOSR Journal of Engineering (IOSRJEN), v. 3, n. 1, p. 05–10, 2013.

RIESENHUBER, M.; POGGIO, T. Riesenhuber, m. poggio, t. hierarchical models of object recognition in cortex. nat. neurosci. 2, 10191025. Nature neuroscience, v. 2, p. 1019–25, 12 1999.

RIESENHUBER, M.; POGGIO, T. How visual cortex recognizes objects: The tale of the standard model (short title: Computational object vision). The Visual Neurosciences, v. 2, 07 2002.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: NAVAB, N. et al. (Ed.). Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. Springer, 2015. (Lecture Notes in Computer Science, v. 9351), p. 234–241. Disponível em: <https://doi.org/10.1007/978-3-319-24574-4_28>.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, v. 65, n. 6, p. 386–408, 1958. ISSN 0033-295X. Disponível em: <http://dx.doi.org/10.1037/h0042519>.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Representations by Back-propagating Errors. Nature, v. 323, n. 6088, p. 533–536, 1986. Disponível em: <http://www.nature.com/articles/323533a0>.

RUSSAKOVSKY, O. et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), v. 115, n. 3, p. 211–252, 2015.

RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. 3. ed. [S.l.]: Prentice Hall, 2010.

SABOUR, S.; FROSST, N.; HINTON, G. E. Dynamic routing between capsules. In: GUYON, I. et al. (Ed.). Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. [s.n.], 2017. p. 3856–3866. Disponível em: <http://papers.nips.cc/paper/6975-dynamic-routing-between-capsules>.

Sam, D. B.; Surya, S.; Babu, R. V. Switching convolutional neural network for crowd counting. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2017. p. 4031–4039.

SEIDLER, J.; MEYER, K.; GILLIVRAY, L. M. Collecting data on crowds and rallies: A new method of stationary sampling. Social Forces, The University of North Carolina Press, v. 55, n. 2, p. 507–519, 1976.

SHEPARD, R. N.; METZLER, J. Mental rotation of three-dimensional objects. Science, American Association for the Advancement of Science, v. 171, n. 3972, p. 701–703, 1971. ISSN 0036-8075. Disponível em: <https://science.sciencemag.org/content/171/3972/701>.

SHI, M. et al. Perspective-aware CNN for crowd counting. CoRR, abs/1807.01989, 2018. Disponível em: <http://arxiv.org/abs/1807.01989>.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556, 09 2014.

SINDAGI, V. A.; PATEL, V. M. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy, August 29 - September 1, 2017. IEEE Computer Society, 2017. p. 1–6. Disponível em: <https://doi.org/10.1109/AVSS.2017.8078491>.

SINDAGI, V. A.; PATEL, V. M. A survey of recent advances in cnn-based single image crowd counting and density estimation. Pattern Recognit. Lett., v. 107, p. 3–16, 2018. Disponível em: <https://doi.org/10.1016/j.patrec.2017.07.007>.

Sindagi, V. A.; Patel, V. M. Ha-ccn: Hierarchical attention-based crowd counting network. IEEE Transactions on Image Processing, v. 29, p. 323–335, 2020.

SU, J.; VARGAS, D. V.; SAKURAI, K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, v. 23, n. 5, p. 828–841, 2019.

VALLOLI, V. K.; MEHTA, K. W-net: Reinforced u-net for density map estimation. CoRR, abs/1903.11249, 2019. Disponível em: <http://arxiv.org/abs/1903.11249>.

WENG, W.; LIN, D. Crowd density estimation based on a modified multicolumn convolutional neural network. In: 2018 IJCNN. [S.l.: s.n.], 2018. p. 1–7. ISSN 2161-4407.

Yan, Z. et al. Perspective-guided convolution networks for crowd counting. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). [S.l.: s.n.], 2019. p. 952–961.

ZEITZ, K. M. et al. Crowd behavior at mass gatherings: a literature review. Prehospital and disaster medicine, Cambridge University Press, v. 24, n. 1, p. 32–38, 2009.

ZENG, L. et al. Multi-scale convolutional neural networks for crowd counting. In: 2017 IEEE International Conference on Image Processing (ICIP). [S.l.: s.n.], 2017. p. 465–469. ISSN 2381-8549.

ZHANG, C. et al. Cross-scene crowd counting via deep convolutional neural networks. In: 2015 IEEE CVPR. [S.l.: s.n.], 2015. p. 833–841.

ZHANG, Y. et al. Single-image crowd counting via multi-column convolutional neural network. In: 2016 IEEE CVPR. [S.l.: s.n.], 2016. p. 589–597. ISSN 1063-6919.