UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

INSTITUTO DE COMPUTAÇÃO - ICOMP

MESTRADO EM INFORMÁTICA - MI

Utilizando redes neurais convolucionais siamesas para filtragem de imagens vazias em dados de armadilhas fotográficas

Luiz Fabio Bailosa de Alencar

Manaus - AM

Fevereiro de 2024

Luiz Fabio Bailosa de Alencar

Utilizando redes neurais convolucionais siamesas para filtragem de imagens vazias em dados de armadilhas fotográficas

Dissertação de mestrado submetida à avaliação, como requisito, para a obtenção do título de Mestre em Informática no Programa de Pós-Graduação em Informática, Instituto de Computação da Universidade Federal do Amazonas.

Orientadora

Eulanda Miranda dos Santos, Ph.D.

Universidade Federal do Amazonas - UFAM

Instituto de Computação - ICOMP

Manaus - AM

Fevereiro de 2024

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Alencar, Luiz Fabio Bailosa de
A368u Utilizando redes neurais convolucionais siamesas para filtragem de imagens vazias em dados de armadilhas fotográficas / Luiz Fabio Bailosa de Alencar . 2024
68 f.: il. color; 31 cm.

Orientadora: Eulanda Miranda dos Santos
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Armadilhas fotográficas. 2. redes siamesas. 3. imagens vazias. 4. rede neurais artificiais. I. Santos, Eulanda Miranda dos. II. Universidade Federal do Amazonas III. Título

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"UTILIZANDO REDES NEURAIS CONVOLUCIONAIS SIAMESAS PARA FILTRAGEM DE IMAGENS VAZIAS EM DADOS DE ARMADILHAS FOTOGRÁFICAS"

LUIZ FABIO BAILOSA DE ALENCAR

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA PELA BANCA EXAMINADORA CONSTITUÍDA PELOS PROFESSORES:

Profa. Dra. Eulanda Miranda dos Santos - PRESIDENTE

Prof. Dr. Luiz Eduardo Soares de Oliveira - MEMBRO EXTERNO

Prof. Dr. Juan Gabriel Colonna - INTERNO

Manaus, 29 de fevereiro de 2024.

Documento assinado eletronicamente por **Eulanda Miranda dos Santos**, **Professor do Magistério Superior**, em 08/06/2024, às 14:46, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Juan Gabriel Colonna**, **Professor**

do Magistério Superior, em 09/06/2024, às 14:44, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Luiz Eduardo Soares de Oliveira**, **Usuário Externo**, em 01/07/2024, às 15:18, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Maria do Perpétuo Socorro Vasconcelos Palheta**, **Secretária em exercício**, em 01/07/2024, às 15:59, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1959001** e o código CRC **BDE95562**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Referência: Processo nº 23105.008641/2024-15 SEI nº 1959001

Dedico este trabalho a todos aqueles a quem esta pesquisa possa ajudar de alguma forma.

Agradecimentos

Primeiramente agradeço aos meus pais, ao grande Luiz Maria Quadros de Alencar e à magnífica Lucineia Oliveira Bailosa por nunca medirem esforços ao me educarem.

Agradeço imensamente à minha querida orientadora professora Eulanda Miranda dos Santos e ao Francisco Fagner do Rego Cunha que seguraram minha mão e me guiaram durante todo processo de pesquisa deste trabalho, e claro, agradecer pela paciência, direcionamentos, sugestões e por me aturarem durante todo esse processo, muito obrigado pessoal! Sem vocês eu não teria chegado tão longe.

Agradeço aos meus grandes amigos Carlos Duarte e Douglas Silva e, em especial, Jéssica Luiza e Rafaela Melo por sempre me ajudarem durante as leituras, atividades, dúvidas, e principalmente, por estarem comigo em momentos de agonia e alegria.

Agradeço às professora Elaine Harada Teixeira,Marcela Pessoa e Fernanda Pires por me incentivarem a ingressar no mestrado, pelos conselhos valiosos, e por me ajudarem durante minha graduação e início da pós-graduação.

Agradeço a todos os professores do PPGI com quem tive contato, especialmente ao professor Eduardo Souto por seus conselhos preciosos, preocupações e revisões.

Por fim, agradeço à UFAM e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) por financiar nossas pesquisas para publicação de artigo pelo programa de AUXÍLIO FINANCEIRO A PROJETO EDUCACIONAL E DE PESQUISA - AUXPE/CAPES.

Utilizando redes neurais convolucionais siamesas para filtragem de imagens vazias em dados de armadilhas fotográficas

Autor: Luiz Fabio Bailosa de Alencar

Orientadora: Eulanda Miranda dos Santos, Ph.D.

Resumo

As câmeras de armadilhas fotográficas são usadas para monitorar a vida selvagem de forma não invasiva através de tarefas como análise populacional de espécies e estudo do comportamento animal ao longo das estações do ano. No entanto, como as imagens são capturadas quando os sensores da câmera detectam movimento, muitas imagens sem animais são capturadas devido ao acionamento dos sensores por outros elementos, como árvores e folhas. Isso resulta em um acúmulo de imagens vazias que ocupam espaço na memória dos equipamentos e consomem largura de banda e energia da rede. Para resolver esse problema, é necessário utilizar métodos que permitam filtrar imagens vazias. No entanto, essa tarefa apresenta desafios, como a variação da vegetação entre diferentes locais, ao longo do dia e das estações do ano. Nesse contexto, o objetivo deste trabalho é desenvolver uma abordagem para filtragem de imagens vazias capturadas por câmeras de armadilhas fotográficas que leve em consideração o ambiente no qual as câmeras estão instaladas. A abordagem proposta é baseada em uma rede neural convolucional siamesa que recebe duas imagens de entrada: 1) uma imagem sem animais que apresenta as características da vegetação local e o nível aproximado de iluminação do dia; e 2) uma imagem capturada a partir do acionamento da câmera que será verificada quanto à presença ou não de animais na

cena. Ao processar as duas imagens, a rede siamesa identifica as diferenças semânticas entre ambas para determinar a existência ou não de animais na imagem capturada. Os resultados obtidos nos experimentos mostram que a abordagem siamesa obteve precisão e acurácia superiores aos resultados obtidos por modelos de classificação que recebem apenas uma imagem por vez, como redes de convolução.

Palavras-chave: Armadilhas fotográficas, redes siamesas, imagens vazias, rede neurais artificiais.

Utilizando redes neurais convolucionais siamesas para filtragem de imagens vazias em dados de armadilhas fotográficas

Autor: Luiz Fabio Bailosa de Alencar

Orientadora: Eulanda Miranda dos Santos, Ph.D.

Abstract

Camera trap images are used to non-invasive wildlife monitoring through tasks such as species population analysis and studying animal behavior throughout the seasons. However, since the images are obtained when the camera's motion sensors are triggered, several images without animals are captured due to the fact that the motion sensors are triggered by other elements, such as trees and leaves. This results in an accumulation of empty images that use memory space and consume bandwidth and network energy. To solve this problem, it is necessary to use methods that allow filtering empty images. However, this a challenging task due to several characteristics, such as the variation of vegetation between different locations and throughout the day and the seasons. In this context, the objective of this work is to present an approach to filter empty images captured by camera trap devices which takes into account information of the environment surrounding the camera. The proposed approach is based on a Siamese convolutional neural network that works with two input images: 1) an image without animals that presents the characteristics of the local vegetation and the approximate level of daylight; and 2) an image captured as usual due to the motion sensor triggering, which will be checked to determine whether or not there are animals in the scene. When processing the two images, the Siamese network identifies the semantic differences

between them so as to identify the presence of animals in the captured image. The obtained results indicate that the Siamese approach reached superior precision and accuracy rates when compared to models that deal with only one image at a time, such as canonical convolutional neural networks.

Keywords: Camera traps, Siamese networks, empty images, artificial neural networks.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1

INTRODUÇÃO

Monitorar a vida selvagem por meio de armadilhas fotográficas é uma alternativa de registro do comportamento dos animais de forma não invasiva (YANG et al., 2021a). Esses equipamentos são posicionados normalmente em áreas de preservação ecológica e capturam uma sequência de fotografias quando algum animal passa pela visão da câmera (HE et al., 2016b). A partir das imagens capturadas, é possível realizar algumas tarefas, tais como analisar o comportamento animal (FREY et al., 2017) e mensurar a quantidade de indivíduos de algumas espécies (RICH et al., 2019). Essa forma de monitoramento permite mapear a vida selvagem durante diferentes períodos do ano e ao longo de vários anos (SCHNEIDER et al., 2019). As câmeras são integradas com sensores que, ao detectarem movimento, são acionados, permitindo a captura de uma pequena sequência de fotografias.

Embora os projetos de armadilhas fotográficas normalmente tenham uma grande quantidade de imagens, a extração de informações dessas imagens é tradicionalmente realizada de forma manual, por meio de especialistas ou grupos de voluntários. Portanto, trata-se de uma tarefa trabalhosa que requer uma grande quantidade de tempo para ser concluída. Por essa razão, há muitas imagens obtidas por câmeras de armadilha fotográfica que permanecem inexploradas (NOROUZZADEH et al., 2018).

Para tentar reduzir esse trabalho manual, existem propostas na literatura para automatizar o processo de classificação de espécies de animais a partir de imagens obtidas por câmeras de armadilha fotográfica. Nessas abordagens, é comum utilizar modelos de CNN (*Convolutional Neural Network*) (BEERY; HORN; PERONA, 2018;

TABAK et al., 2019; WILLI et al., 2019; NOROUZZADEH et al., 2018).

1.1 Motivação e Problema de pesquisa

Um desafio significativo enfrentado pelos sistemas de classificação automática de espécies animais em imagens capturadas por câmeras de armadilhas fotográficas está relacionado à alta quantidade de imagens sem animais registradas pelos sensores dessas câmeras. Isso ocorre principalmente devido às condições ambientais nas quais as câmeras são posicionadas, como a oscilação da vegetação local (WEI et al., 2020). Por exemplo, o sensor da câmera pode ser acionado várias vezes devido ao movimento de objetos que não são animais, como árvores, galhos, folhas e arbustos. Como resultado, a maioria dos conjuntos de dados de imagens de armadilhas fotográficas contém mais imagens vazias do que imagens com a presença de animais. Por exemplo, os conjuntos de dados Snapshot Serengeti (SWANSON et al., 2015) e Elephant Expedition (WILLI et al., 2019) possuem, respectivamente, 75% e 83% de imagens sem a presença de animais.

O acúmulo de imagens vazias é um problema devido a várias razões. Primeiramente, essas são imagens que não agregam informações de valor, uma vez que não possuem animais. Além disso, grande parte dos projetos de armadilhas fotográficas armazena as imagens na memória da câmera, exigindo que a coleta das imagens seja realizada fisicamente, ou seja, os pesquisadores precisam ir até o ponto de captura para efetuar a coleta. Descartar essas imagens pode economizar espaço na memória e, consequentemente, estender o tempo em que a câmera pode ficar fotografando (CUNHA et al., 2021). Evitar o armazenamento desse tipo de imagem também é necessário em sistemas cujo equipamento é conectado à internet, pois economiza largura de banda e energia da rede (ELIAS et al., 2017; CUNHA et al., 2021).

Portanto, torna-se necessária a utilização de métodos que permitam a identificação e o descarte de imagens vazias. Nesse contexto, existem duas principais abordagens empregadas na literatura. A primeira abordagem adiciona a classe de imagem vazia à lista de classes do problema, ou seja, além das classes referentes às espécies de animais, há também a classe de imagem vazia, como é feito em (BEERY; HORN; PERONA, 2018;

TABAK et al., 2019). A segunda abordagem envolve um processo de classificação em duas etapas: 1) um modelo preliminar é utilizado na fase de pré-processamento para identificar as imagens vazias; e 2) um segundo modelo é utilizado para classificar as espécies apenas nas imagens identificadas como não vazias na etapa anterior (WILLI et al., 2019; NOROUZZADEH et al., 2018).

Independentemente da estratégia utilizada para filtrar imagens vazias, um dos principais desafios para os modelos de classificação nesse contexto são as imagens com fundos complexos. As imagens de armadilha fotográfica variam bastante de uma localização para outra. Além disso, mesmo as imagens provenientes do mesmo local podem ser complexas devido à vegetação local, que pode ser densa e ter mudanças rápidas de iluminação em um curto período de tempo (SINGH et al., 2020). Esses fatores podem acabar induzindo os modelos a focarem em aspectos da imagem que não têm relação com os animais (SINGH et al., 2020).

Dessa forma, a filtragem de imagens vazias em bases de imagens de armadilhas fotográficas deve considerar o local de captura e a complexidade do fundo da imagem. Por exemplo, a Figura 1 mostra três imagens que representam desafios comumente encontrados. Na Figura 1(a) e na Figura 1(b), a pelagem do animal possui uma textura semelhante à textura do ambiente, além de existirem muitos detalhes visuais, como galhos, folhas e diferentes tons de iluminação e sombreamento. Embora não tenha relação com o ambiente, algumas imagens são complexas por mostrarem apenas partes do animal, como na Figura 1(c), dificultando a identificação da espécie do animal.

1.2 Justificativa

Estudos mostram que existem imagens de armadilhas fotográficas com fundos visualmente complexos, dificultando a detecção precisa dos animais pelos modelos baseados em CNN (SINGH et al., 2020). A Figura 1 ilustra algumas imagens em que a identificação da espécie do animal presente é difícil, até mesmo para um ser humano. Essa dificuldade ocorre principalmente devido à textura da pele do animal, que apresenta pouco contraste em comparação com as características da vegetação local. Considerando

(a) (b) (c)

Figura 1 – Imagens com fundos complexos: (a) alta variação de iluminação/sombreamento e a textura do animal é semelhante à textura da vegetação; (b) a textura do animal também é semelhante à textura da vegetação; (c) apenas uma parte do animal foi capturada, sendo improvável identificar a espécie. Fonte: (SINGH et al., 2020)

esse contexto, uma possível alternativa para melhorar a precisão dos modelos de filtragem de imagens vazias com fundos complexos seria fornecer aos modelos informações preliminares sobre a região em que a imagem foi capturada.

A arquitetura de rede neural siamesa pode permitir que o modelo compreenda as principais características da região onde as fotografias são capturadas para melhorar a filtragem de imagens vazias. A arquitetura convencional de uma rede neural siamesa é baseada em duas ou mais redes neurais idênticas, com a mesma quantidade de parâmetros. Por exemplo, uma rede neural siamesa composta por dois membros busca aprender representações e/ou relações de duas entradas, ou seja, é capaz de receber duas imagens e calcular suas semelhanças ou diferenças semânticas. Esse tipo de rede é utilizado em diversos problemas de visão computacional, como verificação de rostos (TAIGMAN et al., 2014), rastreamento de objetos (BERTINETTO et al., 2016; CHEN et al., 2020) e validação de assinaturas (DEY et al., 2017).

Uma vez que as redes siamesas são capazes de verificar a similaridade entre imagens e, levando em consideração que as armadilhas fotográficas geram imagens com fundos complexos, a hipótese deste trabalho é a de que utilizar um modelo baseado em CNN siamesa que compare imagens obtidas pelas câmeras com imagens vazias do local de captura, destacando as características da vegetação local, pode ajudar a melhorar a tarefa de filtragem de imagens vazias.

A Figura 2 ilustra o funcionamento esperado com base nessa hipótese. Nesse

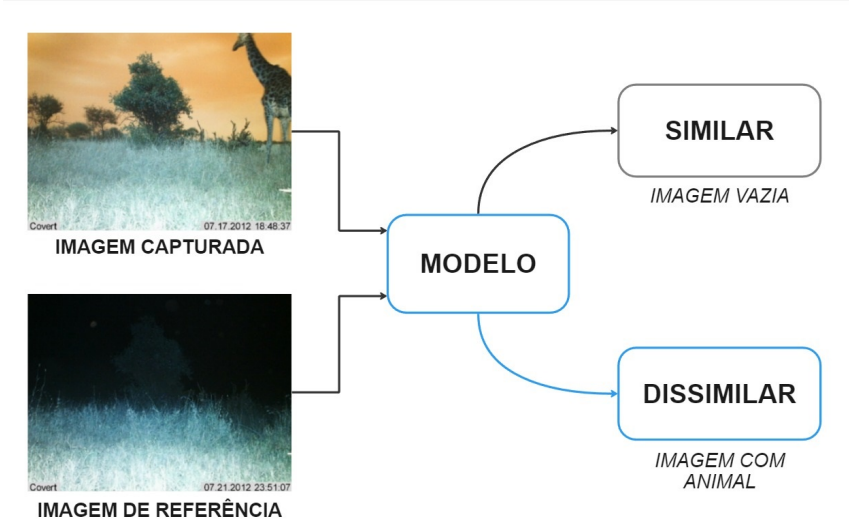


Figura 2 – Processo de comparação via rede siamesa de imagem de referência do local com imagem obtida em processo de captura padrão.

exemplo, são utilizadas redes siamesas compostas por dois membros. Uma imagem vazia do local de captura é designada como "imagem de referência", pois fornecerá à rede siamesa as principais características e elementos complexos do ambiente, como árvores, rochas, galhos e montanhas. Ao comparar a imagem de referência com uma imagem capturada no mesmo local, o modelo calculará a similaridade entre as duas imagens, a fim de determinar se há a presença de um animal na imagem capturada.

Conforme mencionado anteriormente, um dos principais obstáculos para o desenvolvimento de modelos profundos para identificar animais em bases de armadilhas fotográficas é a vegetação dos locais, uma vez que a vegetação varia bastante de um local para outro, mesmo em pontos de captura da mesma região geográfica (AUER et al., 2021). Portanto, utilizar uma imagem de referência que possui informações da vegetação local é importante, visto que ela fornecerá ao modelo aspectos da imagem que podem ser ignorados, destacando apenas a presença do animal.

1.3 Objetivo geral

O objetivo geral deste trabalho é projetar uma abordagem baseada em uma rede neural convolucional siamesa para filtrar imagens de animais em dados de armadilha fotográ-

fica que use uma imagem de referência da vegetação local para reduzir a influência das características do fundo.

1.4 Objetivos específicos

- Elaborar uma estratégia para selecionar as imagens de referência que serão utilizadas para treinar e testar a rede siamesa, pois os pares de imagem devem ser do mesmo ponto de captura.
- Desenvolver e implementar uma arquitetura de rede neural convolucional siamesa para comparar imagens obtidas pelas câmeras com imagens vazias do local de captura.
- Avaliar a eficácia da abordagem proposta em filtrar imagens vazias, comparando os resultados obtidos com outros métodos de filtragem existentes na literatura.

1.5 Contribuições

O método proposto foi publicado no SIBGRAPI 2023 (*Conference on Graphics, Patterns and Images*), e recebeu o prêmio de "*Best Paper*"na trilha principal do evento (ALENCAR; CUNHA; SANTOS, 2023).

A abordagem empregada revelou-se significativamente eficaz na consecução da tarefa complexa de identificação de animais em ambientes naturais. A aplicação de redes siamesas não apenas elevou a acurácia das identificações, mas também delineou perspectivas promissoras para melhorias futuras.

1.6 Organização do Trabalho

Este documento está organizado da seguinte forma: no Capítulo 2 são apresentados alguns conceitos, fundamentos e métricas nos quais a proposta deste trabalho foi baseada. O Capítulo 3 descreve os trabalhos relacionados com o tema desta pesquisa, bem

como uma síntese da literatura. No Capítulo 4 são descritos os detalhes da abordagem proposta para filtragem de imagens vazias. No Capítulo 5 são apresentados resultados do método proposto. Por fim, no Capítulo 6 são apresentadas as considerações finais deste trabalho.

2

FUNDAMENTAÇÃO TEÓRICA

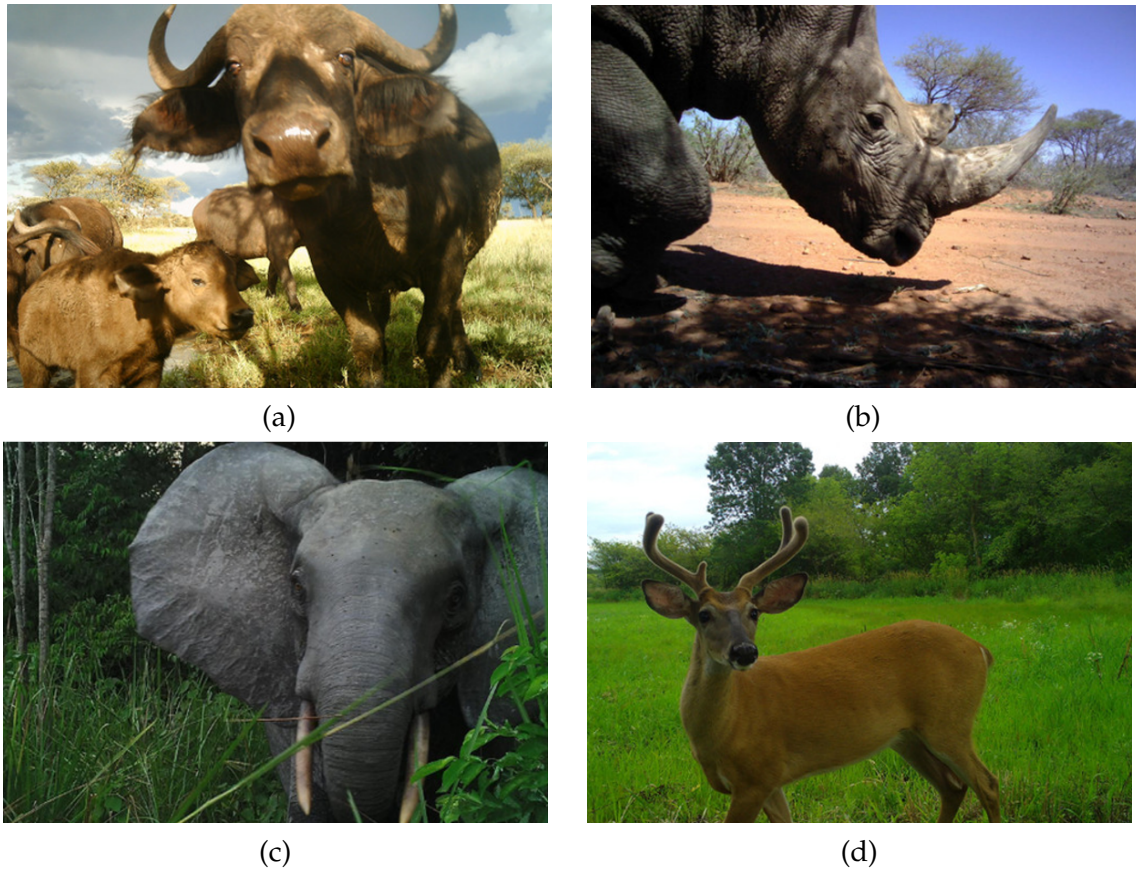
Neste capítulo, são apresentadas informações sobre os principais conceitos para o entendimento da proposta deste trabalho. Inicialmente, são abordados detalhes sobre armadilhas fotográficas e suas aplicações (Seção 2.1). Em seguida, são discutidos definições e conceitos de redes neurais artificiais (Seção 2.2), incluindo sua aplicação em problemas de visão computacional, bem como a estratégia das redes siamesas, que é investigada neste trabalho. Além disso também é apresentado o Zilong (Seção 2.3), uma ferramenta que não utiliza aprendizado de maquina mas pode ajudar na filtragem de imagens vazias em bases de armadilhas fotográficas. Por fim, são apresentadas métricas comumente utilizadas para avaliar modelos de classificação (Seção 2.4).

2.1 Armadilhas fotográficas

Armadilhas fotográficas são câmeras ativadas por calor ou movimento colocadas em ambientes naturais para monitorar e investigar as populações e o comportamento dos animais. Elas são utilizadas para localizar espécies ameaçadas, identificar *habitats* importantes, monitorar locais de interesse e analisar padrões de atividade da vida selvagem (TABAK et al., 2019; NOROUZZADEH et al., 2018). Esses equipamentos são capazes de capturar dezenas de milhares de imagens. No entanto, a extração de informações dessas imagens é tradicionalmente realizada de forma manual. O número de pessoas disponíveis para extrair essas informações é extremamente limitado em comparação com a quantidade de imagens geradas. Por essa razão, grande parte do

valioso conhecimento contido nos repositórios de dados dessas imagens permanece inexplorado (NOROUZZADEH et al., 2018).

Existem várias bases de dados de armadilhas fotográficas disponíveis publicamente. A Figura 3 apresenta imagens de quatro bases de dados públicas: Snapshot Serengeti (a), Camera CATalogue (b), Elephant Expedition (c) e Snapshot Wisconsin (d).



(a) (b)

(c) (d)

Figura 3 – Exemplo de imagens disponíveis em quatro bases de dados públicas: (a) Projeto Snapshot Serengeti; (b) Projeto Camera CATalogue; (c) Projeto Elephant Expedition; e (d) Projeto Snapshot Wisconsin. Fonte: próprio autor.

Os equipamentos das armadilhas fotográficas são normalmente posicionados em localizações estratégicas. Por exemplo, a Figura 4 mostra a área de estudo do projeto Snapshot Serengeti, indicada pela linha tracejada, bem como a localização e distribuição das câmeras do referido projeto. Os equipamentos normalmente são posicionados em árvores e podem também ser fixados em postes de aço, conforme mostra a Figura 5.

A automação do processo de análise dessas imagens é um campo de pesquisa amplamente investigado. As redes neurais têm sido massivamente utilizadas para enfrentar esse desafio, como demonstrado em estudos anteriores (WILLI et al., 2019;

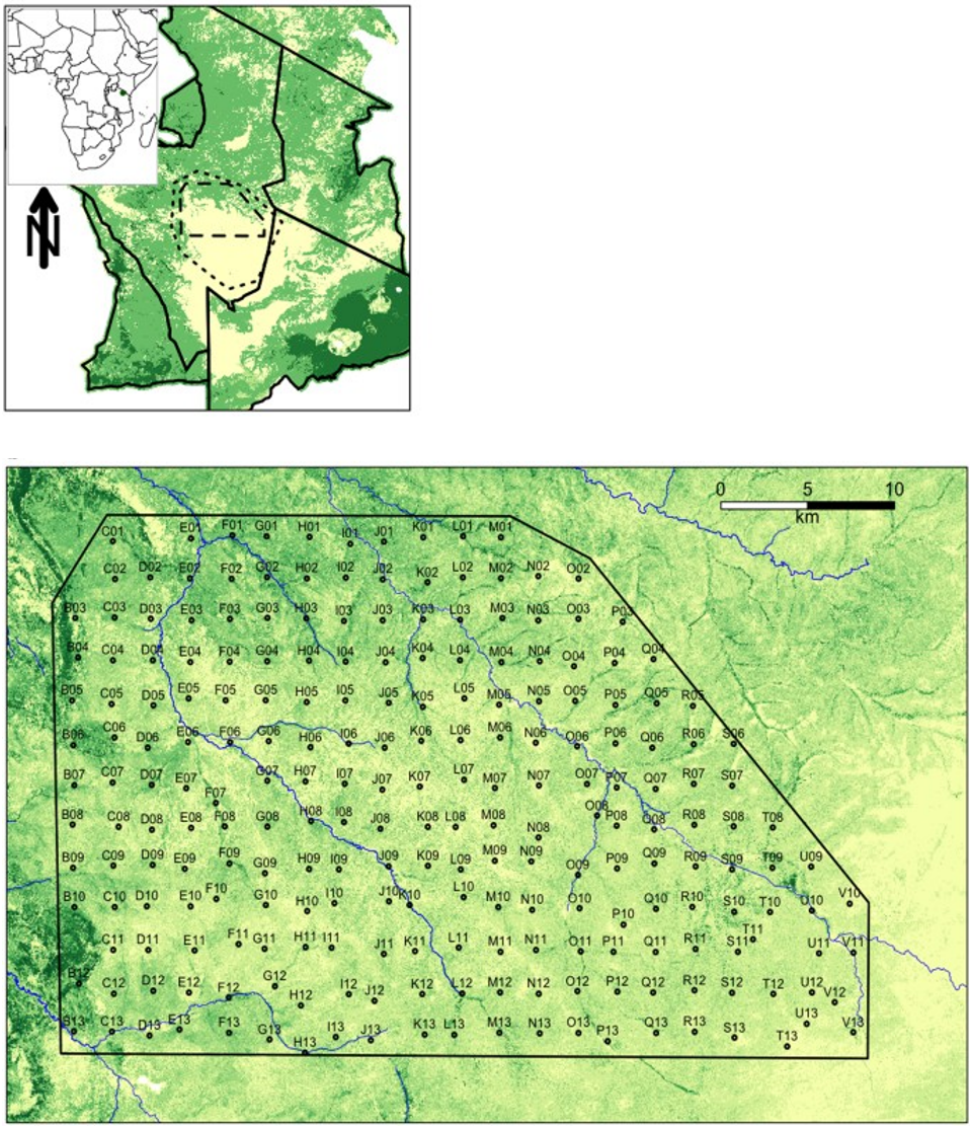


Figura 4 – Localização de câmeras do projeto Snapshot Serengeti; a área de estudo do projeto é indicada pela linha tracejada. Fonte: (SWANSON et al., 2015).

NOROUZZADEH et al., 2018; CUNHA et al., 2021; YANG et al., 2021b; TABAK et al., 2019; WEI et al., 2020; YANG et al., 2021c; SWANSON et al., 2015). No entanto, a análise dessas imagens apresenta diversos desafios, tais como problemas na qualidade das imagens, variações na iluminação ao longo do dia e uma alta taxa de imagens sem a presença de animais. Esse último aspecto é principalmente causado por animais que se movem rapidamente, dificultando a captura das imagens, e também por áreas onde a vegetação do ambiente, como árvores, está constantemente em movimento, resultando em capturas de imagens sem a presença de animais devido aos sensores de movimento da câmera (WEI et al., 2020).

Figura 5 – Armadilha fotográfica do projeto Snapshot Serengeti. Fonte: (SWANSON et al., 2015).

2.1.1 Desafios na filtragem de imagens em bases de armadilhas fotográficas

Conforme mencionado anteriormente, um dos principais desafios relacionados às imagens capturadas por armadilhas fotográficas é a alta proporção de imagens vazias, ou seja, sem a presença de animais. Por exemplo, Swanson et al. (2015) demonstraram que aproximadamente 75% das imagens na base de dados Snapshot Serengeti são classificadas como vazias. Essa predominância de imagens sem animais dificulta a extração de informações relevantes sobre as espécies.

Outro desafio decorrente da grande quantidade de imagens vazias está relacionado ao desbalanceamento de classes. Dependendo da localização em que as câmeras são instaladas, ocorre naturalmente um desequilíbrio entre as espécies. O desbalanceamento entre as classes é um problema desafiador para algoritmos de aprendizado de máquina, pois eles tendem a se especializar na classificação das classes majoritárias. Esse problema é ainda mais agravado pela presença massiva de imagens vazias, como pode ser observado na Figura 6, que apresenta a distribuição da quantidade de imagens para as 15 classes mais representadas na base de dados Serengeti. É possível observar

que a classe "empty" corresponde a 75% das instâncias, enquanto a espécie com maior representação de imagens é a classe "wildebeest", com apenas 10%.

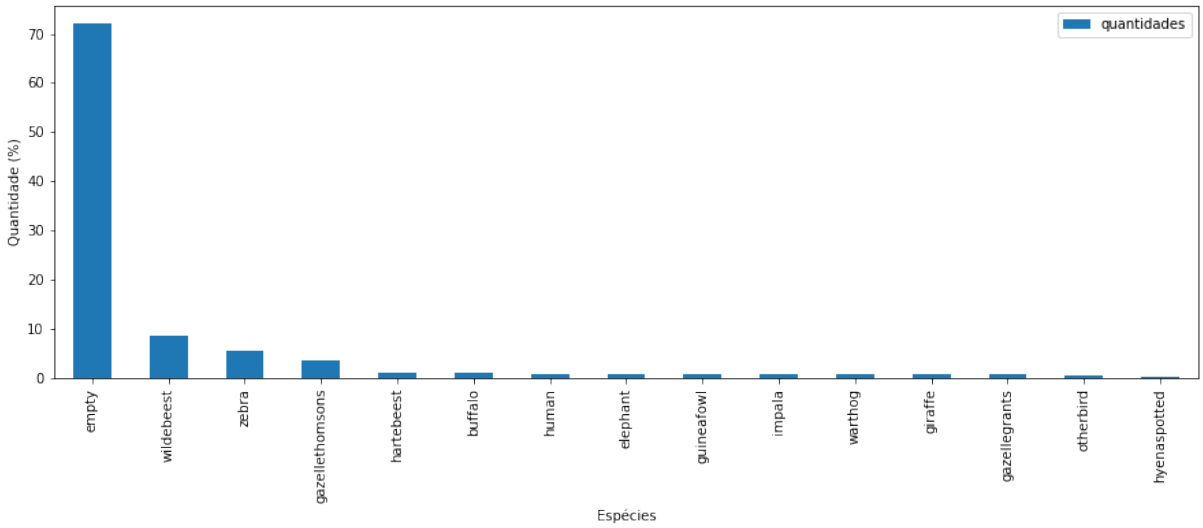


Figura 6 – Distribuição das espécies de animais na base de dados Snapshot Serengeti. Fonte: próprio autor.

2.2 Redes neurais artificiais

Na ciência da computação e áreas afins, as redes neurais artificiais (RNAs) são modelos computacionais inspirados no sistema nervoso de animais e são capazes de aprender características através do reconhecimento de padrões (GOODFELLOW; BENGIO; COURVILLE, 2016). RNAs são frequentemente caracterizadas como sistemas de sensores conectados e tentam imitar o comportamento das redes neurais biológicas. Os modelos de aprendizado de máquina baseados em RNAs são utilizados em diversos problemas do mundo real, como identificação de doenças (GOLHANI et al., 2018), detecção de objetos (DHILLON; VERMA, 2020) e previsão de séries temporais (TEALAB, 2018), além de contribuir para a automatização de processos em áreas como produção industrial e exploração de petróleo (ABIODUN et al., 2018).

Os principais problemas de aprendizado de máquina em que as RNAs são aplicadas podem ser agrupados em problemas de classificação e de regressão. Um problema de classificação simples é baseado na identificação de uma classe para cada

instância do problema, dentre um conjunto de n classes possíveis. Quando n é igual a 2, o problema também pode ser chamado de classificação binária (REZAEI; LIU, 2019). Nos problemas de regressão, as redes buscam prever um valor numérico contínuo, como, por exemplo, prever a temperatura climática de uma região ou a idade de uma pessoa (PENG et al., 2021).

O processo de aprendizagem das RNAs é realizado através de uma etapa de treinamento. Durante essa etapa, as redes buscam aprender características de um problema através de instâncias de treino. Em problemas de classificação, é gerada uma função $f(x) = y$, sendo f uma função gerada por um modelo, x uma instância com um determinado número de atributos e y o rótulo previsto pelo modelo (GOODFELLOW; BENGIO; COURVILLE, 2016). A função gerada do modelo é proveniente de um conjunto de pesos que são ajustados durante a etapa de treinamento e servem para identificar quais são os principais atributos de uma instância, bem como a relação que os valores dos atributos possuem entre si. A Figura 7 ilustra um modelo classificador de maçãs, no qual x é a imagem de entrada cujos atributos são as cores e composição dos pixels da imagem, $f(x)$ é o modelo classificador, e y é a predição do modelo.

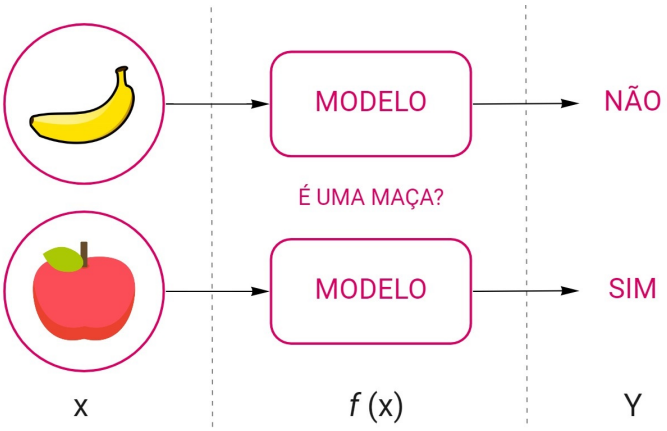


Figura 7 – Problema de classificação de maçãs. Fonte: Próprio autor.

A arquitetura de uma RNA, ilustrada na Figura 8, é composta por várias camadas, sendo que cada camada pode possuir diferentes neurônios que estão conectados por meio pesos. A camada que recebe os atributos de uma instância é conhecida como camada de entrada; a previsão do modelo é o resultado da camada de saída; e as camadas internas são conhecidas como camadas escondidas. Dentre os diversos mo-

delos de RNAs disponíveis na literatura, neste trabalho serão utilizadas redes neurais convolucionais, descritas na próxima seção.

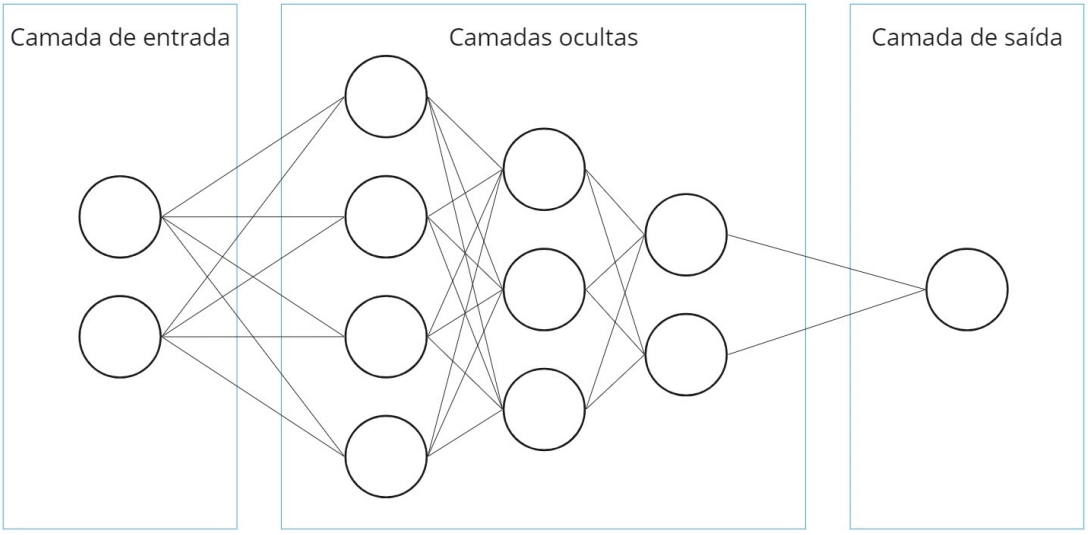


Figura 8 – Exemplo da arquitetura de uma rede neural artificial. Fonte: próprio autor.

2.2.1 Redes neurais convolucionais

A visão computacional é um dos ramos da inteligência artificial que investiga estratégias de processamento de imagens e vídeos, fornecendo ao computador a capacidade de lidar com problemas visuais (GOODFELLOW; BENGIO; COURVILLE, 2016). As arquiteturas de RNAs normalmente aplicadas em problemas de visão computacional são as redes neurais convolucionais, do inglês *Convolutional Neural Network* - CNN. Essas redes conseguem tratar uma variedade de problemas envolvendo imagens e vídeos, como detecção de objetos (OUYANG et al., 2016; DIBA et al., 2017), rastreamento de movimento (DOULAMIS; VOULODIMOS, 2016; DOULAMIS, 2018), reconhecimento de ações (LIN et al., 2016; CAO; NEVATIA, 2016), estimativa de pose humana (TOSHEV; SZEGEDY, 2014; CHEN; YUILLE, 2014) e segmentação semântica (NOH; HONG; HAN, 2015; LONG; SHELHAMER; DARRELL, 2015).

As CNNs processam os dados nas camadas de entrada e camadas escondidas utilizando uma topologia semelhante a uma grade (Figura 9), ajudando o modelo a capturar dependências temporais e espaciais de uma imagem através da aplicação

de filtros em camadas de convolução. As camadas de convolução são as principais propriedades que diferenciam as arquiteturas de CNNs entre si, pois essas camadas servem para mapear as principais características de uma imagem. Em um problema de classificação de animais, por exemplo, os pesos dessas camadas são essenciais para o modelo entender as características de cada espécie de animal.

As camadas de convolução podem estar conectadas por outras camadas do mesmo tipo, assim como camadas de *pooling*, que são utilizadas para reduzir as dimensões do mapa de características de camadas anteriores. A Figura 9 ilustra uma arquitetura CNN com múltiplas camadas de convolução e uma camada de *pooling* que vai afunilando as características da camada de entrada até à camada de saída para calcular a probabilidade de a imagem de entrada pertencer à classe "gato".

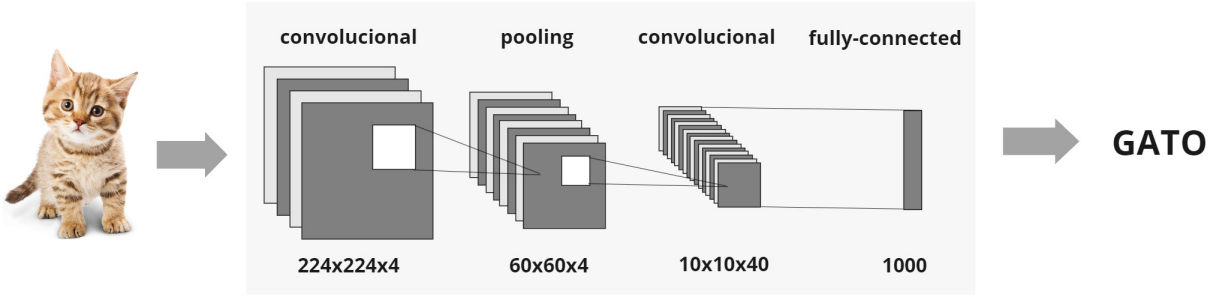


Figura 9 – Exemplo de arquitetura de uma CNN com camadas de convolução e de *pooling*. Fonte: próprio autor.

2.2.2 MobileNetV2

A MobileNetV2 (SANDLER et al., 2018) é uma CNN que é utilizada neste trabalho, ela foi projetada para oferecer alto desempenho em dispositivos móveis com recursos computacionais limitados. Ela se destaca por sua eficiência computacional e de memória, sem sacrificar a acurácia em tarefas de visão computacional.

A MobileNetV2 baseia-se na arquitetura da MobileNet original, mas com diversos aprimoramentos:

- Blocos Inverted Residual: A principal inovação da MobileNetV2. Esses blocos invertem a ordem tradicional das operações em um bloco residual, começando

com uma expansão de canais e terminando com uma redução de canais. Isso permite uma melhor relação entre eficiência e desempenho.

- Escala de Canais Dinâmicos: A MobileNetV2 utiliza um método de escala de canais dinâmicos para ajustar automaticamente o número de canais em cada bloco residual, de acordo com a dificuldade da tarefa. Essa estratégia otimiza o uso de recursos computacionais.

- Linear Bottleneck: Uma camada de linear bottleneck é utilizada entre os blocos inverted residual para reduzir a dimensionalidade das features, antes da operação de expansão de canais. Isso contribui para a eficiência da rede.

As principais vantagens da arquitetura MobileNetV2 englobam sua adequação para dispositivos móveis com recursos computacionais restritos, sua acurácia comparável a outras CNNs em tarefas de visão computacional e sua adaptabilidade fácil para diferentes finalidades, tais como classificação de imagens, detecção de objetos e segmentação semântica.

Entre as desvantagens identificadas, destaca-se a potencial complexidade da arquitetura da MobileNetV2 em comparação com outras CNNs, o que pode dificultar tanto sua interpretação quanto seu ajuste fino. Além disso, é importante notar que a MobileNetV2 foi inicialmente concebida para a tarefa específica de classificação de imagens, o que pode limitar sua eficácia em outras aplicações.

2.2.3 ResNet50

A ResNet50 (HE et al., 2016a) é uma outra CNN de última geração que é utilizada neste trabalho, ela é conhecida por sua alta acurácia em tarefas de visão computacional, especialmente em classificação de imagens. Ela se destaca por sua arquitetura inovadora que utiliza "atalhos"para aliviar o problema do desvanecimento do gradiente durante o treinamento.

A ResNet50 é composta por uma série de blocos residuais, cada um consistindo de:

- Camadas convolucionais: Extraem features da imagem.
- Função de ativação: ReLU é a função mais utilizada.
- Camadas de normalização: Batch normalization é utilizada para estabilizar o treinamento.

Os blocos residuais são conectados por "atalhos", que permitem que as features de camadas anteriores sejam propagadas diretamente para as camadas posteriores. Isso ajuda a evitar o problema do desvanecimento do gradiente e facilita o treinamento de redes mais profundas. A ResNet50 é mais complexa e computacionalmente cara do que outras CNNs, como a VGGNet (SIMONYAN; ZISSERMAN, 2014). Isso se deve ao uso de um grande número de camadas e à presença de "atalhos".

As vantagens de empregar a rede ResNet50 residem no fato de que esta se destaca como uma das CNNs mais precisas para a classificação de imagens. Sua arquitetura profunda permite a aprendizagem de características complexas pela rede, tornando-a versátil o suficiente para ser facilmente adaptada para uma variedade de tarefas, como detecção de objetos e segmentação semântica.

Por outro lado, as desvantagens do seu uso são decorrentes da complexidade intrínseca da arquitetura, o que pode dificultar tanto a interpretação dos resultados quanto o refinamento fino da rede. Além disso, é importante destacar que a ResNet50 demanda mais recursos computacionais em comparação com outras CNNs, o que pode tornar sua implementação mais onerosa em termos de custo computacional.

2.2.4 EfficientNetB0

A EfficientNetB0 (TAN; LE, 2019) é uma CNN bem conhecida que também é utilizada neste trabalho, ela foi projetada para oferecer alto desempenho com baixo custo computacional. Ela se destaca por sua arquitetura eficiente que combina diferentes escalas de features e otimiza o uso de recursos computacionais.

A EfficientNetB0 baseia-se na arquitetura da EfficientNet original, mas com algumas modificações:

- Escala de Canais Dinâmicos: A EfficientNetB0 utiliza um método de escala de canais dinâmicos para ajustar automaticamente o número de canais em cada bloco da rede, de acordo com a dificuldade da tarefa. Isso otimiza o uso de recursos computacionais.

- Compound Scaling: A EfficientNetB0 utiliza um método de "compound scaling"para aumentar a capacidade da rede de forma eficiente. Isso envolve aumentar a resolução da imagem de entrada, a profundidade da rede e o número de canais de forma proporcional.

- Blocos MBConv: A EfficientNetB0 utiliza blocos MBConv, que são uma combinação de blocos MobileNetV2 e Inverted Residual.

A EfficientNetB0 é significativamente mais eficiente que outras CNNs comparáveis, como a ResNet50. Isso se deve à sua arquitetura otimizada e à utilização de técnicas de escala de canais dinâmicos e "compound scaling".

As principais vantagens da EfficientNetB0 incluem sua aptidão para ambientes com recursos computacionais restritos, como dispositivos móveis e plataformas com capacidades limitadas. Além disso, demonstra uma acurácia comparável a outras CNNs em tarefas de visão computacional e apresenta uma notável adaptabilidade para diversas finalidades, como classificação de imagens, detecção de objetos e segmentação semântica.

Entre as desvantagens, observa-se uma analogia com a MobileNetV2. A arquitetura da EfficientNetB0 é relativamente mais complexa em comparação com outras CNNs, o que pode acarretar dificuldades na interpretação e no ajuste fino dos modelos. Além disso, é válido ressaltar que a EfficientNetB0 foi originalmente concebida para a classificação de imagens, o que pode limitar sua eficácia em outras tarefas.

2.2.5 Redes siamesas

As redes siamesas são modelos que possuem dois ou mais submodelos internos que são idênticos e normalmente compartilham os mesmos pesos (HE et al., 2018). Essa

abordagem busca tratar problemas envolvendo a similaridade de duas ou mais entradas, sendo utilizada frequentemente em aplicações de verificação de rostos (TAIGMAN et al., 2014; SONG et al., 2019) e rastreamento de objetos (BERTINETTO et al., 2016; CHEN et al., 2020).

A proposta principal é verificar se as entradas são semelhantes. Por exemplo, em um problema de verificação de assinaturas o objetivo é checar se duas assinaturas são do mesmo cliente, conforme ilustrado na Figura 10. Nesse exemplo, cada assinatura é processada paralelamente por uma das duas CNNs internas. Em seguida, cada CNN gera uma saída: $s1$ e $s2$, que correspondem respectivamente às características da primeira e da segunda entrada. Por fim, é realizado um cálculo de distância L, como a euclidiana, entre $s1$ e $s2$, gerando a previsão do rótulo y.

Em problemas supervisionados, tradicionalmente as redes neurais são treinadas para prever um determinado número de classes. Porém, caso haja necessidade de remover ou adicionar novas classes, deve ser realizada uma nova rodada de treinamento para o modelo considerar novas classes. A arquitetura das redes siamesas, por outro lado, permite que o modelo aprenda uma função de similaridade L, evitando que o modelo treine novamente para aprender novas classes.

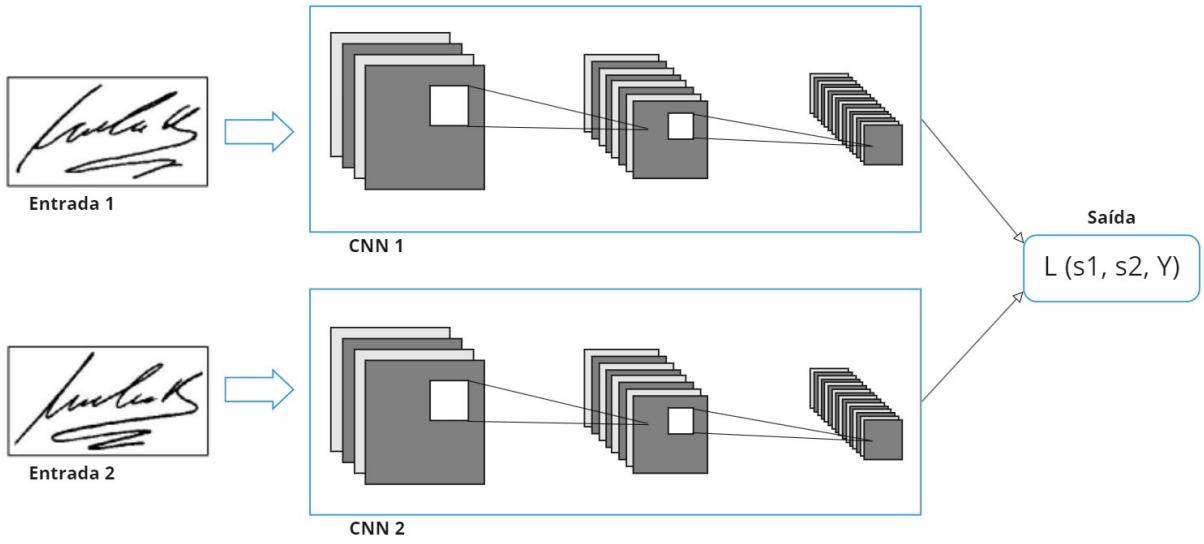


Figura 10 – Exemplo de uma rede neural artificial com arquitetura siamesa. Fonte: próprio autor.

Além de problemas de classificação binária, as redes siamesas também são

utilizadas em aplicações para detecção de mudanças em imagens. A SNUNet-CD (FANG et al., 2021), por exemplo, é uma rede cuja arquitetura é baseada na rede U-Net (RONNEBERGER; FISCHER; BROX, 2015). Essa rede recebe duas entradas, semelhante às redes siamesas tradicionais, mas sua saída é uma máscara que representa a diferença semântica entre as duas entradas.

2.3 Zilong

Nesta seção é apresentado também o Zilong, que é uma ferramenta de software gratuita que utiliza um algoritmo de aprendizado não-supervisionado para detectar imagens vazias em dados coletados por câmeras de armadilhas (WEI et al., 2020). Esse método é um *baseline* importante utilizado neste projeto de mestrado e por essa razão é descrito aqui. Trata-se de uma ferramenta de linha de comando, construída no ambiente R, compatível tanto com sistemas Windows quanto Linux. Uma característica importante do Zilong é a sua independência de dependências adicionais durante a execução, além da habilidade de preservar a maioria das imagens e eventos com animais, o que o torna ideal para filtrar imagens irrelevantes antes de empregar modelos de aprendizado profundo na classificação de espécies de animais (WEI et al., 2020).

Ao realizar a identificação de imagens sem conteúdo relevante (imagens vazias), o Zilong analisa os dados provenientes das câmeras de armadilhas, distinguindo entre imagens com animais e aquelas sem. A ferramenta não tem o propósito de identificar as espécies de animais, concentrando-se exclusivamente na filtragem das imagens irrelevantes, o que pode prejudicar o desempenho dos modelos de treinamento. Dessa forma, o Zilong se apresenta como uma solução eficiente para economizar tempo e recursos dos pesquisadores da vida selvagem, ao evitar a revisão manual de todas as imagens coletadas pelas armadilhas fotográficas, proporcionando assim uma forma econômica de processar essas imagens(WEI et al., 2020).

A implementação do Zilong assume que, se alguma atividade acionar a câmera, uma sequência de imagens será gravada em um curto período de tempo e essas imagens compartilharão o mesmo fundo. Ao comparar a diferença de valor de pixel e mudança

de pixels entre essas imagens, Zilong pode distinguir entre imagens de animais e imagens vazias. O algoritmo é baseado na suposição de que o fundo permanece o mesmo em uma sequência de imagens, enquanto a presença de animais causa mudanças nos valores dos pixels.

O Zilong está disponível gratuitamente sob a licença BSD (*Berkeley Software Distribution*) e acessível no GitHub [1]. Trata-se de uma ferramenta valiosa para os pesquisadores da vida selvagem, permitindo revisar as imagens das armadilhas fotográficas de forma econômica, sem a necessidade de recorrer a abordagens computacionalmente intensivas de aprendizado de máquina.

2.4 Métricas de avaliação

Uma forma simples de avaliar predições de modelos em problemas de classificação é através da matriz de confusão. Nessa abordagem, é possível identificar a quantidade de predições corretas e incorretas em cada classe (Figura 11). Em um problema de classificação binária temos a classe “positiva” e a classe “negativa”. Na matriz de confusão, as instâncias das classes previstas em relação ao valor real são organizadas nos seguintes grupos: TP, FN, FP e TN.

- **TP:** corresponde à quantidade de instâncias da classe real “positiva"que foram classificadas como “positiva";
- **FN:** corresponde à quantidade de instâncias da classe real “positiva” que foram classificadas como “negativa”;
- **FP:** corresponde à quantidade de instâncias da classe real “negativa” que foram classificadas como “positiva”; e
- **TN:** corresponde à quantidade de instâncias da classe real “negativa” que foram corretamente classificadas como “negativa”.

[1] Disponível em: https://github.com/0106WeiWeiDeng/Zilong

Real	Previsão		
		Positivo	Negativo
	Positivo	TP	FN
	Negativo	FP	TN

Figura 11 – Matriz de confusão de um problema de classificação binária. Fonte: Próprio autor.

Por meio da matriz de confusão é possível calcular a acurácia do modelo. A acurácia é uma métrica que avalia a porcentagem das instâncias previstas corretamente independentemente da classe. A acurácia é calculada da seguinte maneira:

$$\text{Acurácia} = \frac{TP+TN}{TP+FP+FN+TN}$$

De forma geral, a acurácia consegue apresentar um resultado superficial do desempenho do modelo. Como desvantagem, entretanto, essa métrica não avalia cada classe de maneira independente. Logo, atingir uma acurácia alta não significa que o modelo está classificando bem as instâncias.

A precisão é outra métrica de avaliação que utiliza os resultados da matriz de confusão. A precisão consegue avaliar o desempenho do modelo em cada uma das classes e indica a porcentagem de acertos para uma determinada classe, diferentemente da acurácia. Para medir a precisão da classe positiva, temos a seguinte formula:

$$\text{Precisão} = \frac{TP}{TP+FP}$$

A revocação também avalia as classes independentemente. Semelhantemente à precisão, a revocação mostra a razão de instâncias de uma classe que foram corretamente previstas através da fórmula:

$$\text{Revocação} = \frac{TP}{TP+FN}$$

Por fim, outra métrica popular é o F1-Score, que leva em consideração os valores de precisão e revocação calculando uma média harmônica entra as duas métricas. A formula do F1-Score é definida abaixo. Essa métrica apresenta um resumo mais preciso da qualidade do modelo, pois seu valor será alto somente se precisão e revocação forem elevadas também.

$$F1 = 2 * \frac{\text{Revocação} * \text{Precisão}}{\text{Revocação} + \text{Precisão}}$$

Existem métricas também que tentam medir os tipos de erro dos modelos predi-

tores como o erro de comissão e omissão. Erro de comissão é definido como a fração de valores que foram previstos para estar em uma classe, mas que não pertencem a essa classe, ou seja, os erros de comissão representam falsos positivos. Erro de omissão é definido como a fração de valores que pertencem a uma classe, mas foram previstos para estar em uma classe diferente, em outras palavras, os erros de omissão representam falsos negativos. Ambos os tipos de erros podem ser calculados através das fórmulas abaixo:

$$\text{Erro de omissão} = \frac{FN}{FN+TP}$$

$$\text{Erro de comissão} = \frac{FN}{FN+TN}$$

2.5 Síntese do capítulo

Este capítulo abordou diversos tópicos relacionados à proposta deste trabalho, incluindo conceitos fundamentais sobre armadilhas fotográficas e suas aplicações, definições e conceitos de redes neurais artificiais, a estratégia das redes siamesas, a apresentação do Zilong (uma ferramenta de software para filtragem de imagens vazias em dados de armadilhas fotográficas), e métricas comumente utilizadas para avaliar modelos de classificação.

As armadilhas fotográficas são descritas como câmeras utilizadas para monitorar e investigar a vida selvagem em ambientes naturais, com aplicações que vão desde a localização de espécies ameaçadas até a análise de padrões de atividade dos animais. No entanto, o processamento manual das imagens capturadas por essas câmeras é inviável devido ao grande volume de dados, resultando na subutilização do conhecimento contido nessas imagens.

A automação desse processo de análise é um campo de pesquisa em expansão, com as redes neurais, em particular as redes neurais convolucionais, desempenhando um papel crucial. No entanto, a análise dessas imagens enfrenta desafios como a alta proporção de imagens vazias e o desbalanceamento de classes, o que pode prejudicar o desempenho dos modelos de aprendizado de máquina.

As redes siamesas são introduzidas como uma abordagem para tratar proble-

mas de similaridade entre duas ou mais entradas, sendo aplicadas em diversas áreas, como verificação de rostos e detecção de mudanças em imagens. O Zilong é apresentado como uma ferramenta de software que utiliza um algoritmo de aprendizado não supervisionado para detectar imagens vazias em dados de armadilhas fotográficas, proporcionando uma solução eficiente para economizar tempo e recursos dos pesquisadores.

Além disso, este capítulo descutiu métricas comumente utilizadas para avaliar modelos de classificação, incluindo a matriz de confusão, acurácia, precisão, revocação e F1-Score, bem como os conceitos de erro de comissão e erro de omissão. Essas métricas fornecem uma avaliação detalhada do desempenho dos modelos e ajudam a identificar possíveis áreas de melhoria.

3

TRABALHOS RELACIONADOS

A classificação de imagens vazias em armadilhas fotográficas normalmente é tratada em duas principais abordagens: 1) considerar as imagens vazias no treinamento do modelo como uma classe adicional, além das classes das espécies dos animais (BEERY; HORN; PERONA, 2018; SCHNEIDER et al., 2020); ou, 2) realizar um filtro das imagens sem animais na etapa de pré-processamento, utilizando um classificador binário para determinar a existência ou não de animais nas imagens (NOROUZZADEH et al., 2018; WILLI et al., 2019). Nesse caso, o modelo classificador de espécies utiliza apenas imagens de animais. Este capítulo apresenta trabalhos com métodos baseados na segunda abordagem, isto é, um problema de classificação binária de imagens vazias na etapa de pré-processamento.

3.1 Classificação de imagens vazias em armadilhas fotográficas utilizando aprendizado profundo

No trabalho de Norouzzadeh et al. (2018), os autores investigaram três diferentes tipos de tarefa: classificar imagens, contar indivíduos e descrever as imagens de animais utilizando aprendizado profundo. Embora a parte de classificação tenha sido focada nas espécies dos animais, foi realizado um experimento para verificar o desempenho de vários modelos ao filtrar imagens vazias de armadilhas fotográficas. Os modelos utilizados foram: AlexNet, NiN, VGG, GoogLeNet e ResNet. Durante a etapa de treinamento

foi utilizada uma base de dados balanceada de 1.4 milhões de imagens (757.000 imagens vazias e 757.000 imagens de animais) e na etapa de teste 105.000 imagens. Todas as arquiteturas alcançaram acurácia superior a 95.8%, sendo o VGG o modelo com maior acurácia, atingindo 96.8%.

Willi et al. (2019) também avaliaram o desempenho de várias CNNs para classificar imagens em diferentes espécies de animais, imagens vazias e imagens de humanos ou veículo. Novamente, embora o estudo tenha sido concentrado na classificação das espécies, os autores também avaliaram o impacto do uso da técnica de transferência de aprendizagem na tarefa de filtragem de imagens vazias. Os autores utilizaram a arquitetura ResNet18 em duas rodadas de treinamento em 4 bases de dados de armadilhas fotográficas: Snapshot Serengeti, Câmera CATalogue , Elephant Expedition e Snapshot Wisconsin. Uma das rodadas foi realizada utilizando transferência de aprendizagem e ajuste fino do modelo na base Snapshot Serengeti, enquanto na outra rodada não foi feito ajuste fino nessa base. A precisão na filtragem de imagens vazias entre as bases de dados variou de 91,2% a 98,0%.

Considerando que o desbalanceamento dos dados é um dos principais problemas gerados pelas imagens vazias, pois a quantidade dessas imagens geralmente é maior do que a quantidade de imagens com presença de animais, Yang et al. (2021a) estudaram o impacto desse problema ao variar as proporções de imagens vazias (EIR - *Empty Image Rate*) da base de dados Snapshot Serengeti (de 10% a 90% de proporção). Os autores utilizaram o modelo AlexNet em seus experimentos. Os resultados mostraram que quando a proporção das imagens vazias está entre 10% e 70%, o desbalanceamento apresenta pouco efeito no desempenho do modelo. Porém, o estudo também indicou que os erros de omissão e de comissão podem ser ajustados de acordo com a proporção de imagens vazias utilizadas durante o treinamento: geralmente quanto mais equilibrada a classe vazia está em relação à classe de animal, menores são os erros de omissão e comissão. Por exemplo, quando o EIR foi definido em 40%, os erros de comissão e de omissão foram 7.4% e 7.3% respectivamente, mas, ao mudar o EIR para 80%, os erros de comissão e de omissão aumentaram para 12.5% e 8.7% respectivamente.

Para tentar tratar o problema de desbalanceamento das imagens vazias, Yang et

al. (2021c) utilizaram uma abordagem de conjunto de preditores com três modelos de CNN: AlexNet, ResNet e Inception. Os autores realizaram o treinamento utilizando a base de dados Lhasa Mountain em duas etapas: com a base de dados desbalanceada; e, preservando os pesos do treinamento anterior, com a base de dados balanceada. Nesse estudo, os modelos individuais de CNN utilizaram *thresholds*[1] de 50%, 90%, 92.5%, 95% e 97.5%, para determinar os votos do conjunto de preditores. Os autores indicaram que os resultados são promissores. Ao utilizar *threshold* de 95%, a combinação dos preditores obteve erro de omissão de 0.7%. Contudo, pelo motivo dos modelos individuais utilizarem um valor de *threshold* alto, o comitê conseguiu identificar apenas 47.66% das imagens da base de dados.

Seguindo uma abordagem diferente do tratamento supervisionado realizado nos trabalhos descritos anteriormente, Yang et al. (2021b) propuseram um método de treinamento incremental para filtrar imagens vazias de armadilhas fotográficas. Dado que o desenvolvimento de um modelo robusto normalmente exige uma grande quantidade de instâncias de treinamento para conseguir generalizar bem na aplicação real, esses autores aplicaram um treinamento incremental utilizando poucas imagens para gerar um modelo classificador. A abordagem utilizou o modelo AlexNet e a etapa de treinamento foi fragmentada em 7 rodadas sequenciais de treino. Cada rodada de treinamento aumentou o número de instâncias empregadas, mas os pesos do modelo foram reaproveitados. A proporção das imagens vazias em todos os conjuntos de treino foi de 80%. O modelo foi testado em um conjunto de 7.298 instâncias, atingindo erro geral, de comissão e de omissão igual a 2.69%, 6.82% e 6.45% respectivamente.

Existem cenários em que a filtragem das imagens vazias precisa ocorrer no momento em que a fotografia é capturada para evitar armazenar informações desnecessárias. Nesse sentido, Cunha et al. (2021) realizaram um estudo comparativo do desempenho de classificadores e detectores cujas arquiteturas podem ser utilizadas em sistemas embarcados, ou seja, podem filtrar a imagem no momento da captura. Os modelos utilizados para classificar são: MobileNetV2, EfficientNetB0 e EfficientNetB3; e para a detecção: SSDLite+MobileNetV2 e EfficientDet-D0. Para comparar os resultados dos modelos foram utilizadas as bases de dados Caltech e Serengeti. A partir dos resul-

[1] Threshold: confiança que um modelo possui sobre as classes previstas

tados foi constatado que o EfficientDet-D0 foi capaz de eliminar mais do que o dobro de imagens vazias quando comparado aos outros classificadores usando a base de dados Caltech. Como esperado, os detectores também obtiveram resultados melhores do que os classificadores usando a base de dados Serengeti, com uma margem de 8% de precisão em que o modelo MobileNetV2 atingiu 82.89% e o EfficientNet-B3 com 87.67%. Apesar do desempenho dos detectores ser superior, sua latência de inferência também é maior e pode limitar seu uso em dispositivos de borda (CUNHA et al., 2021). Além disso, os classificadores podem obter resultados satisfatórios, caso o conjunto de dados de treinamento seja grande.

3.2 Classificação de imagens vazias em armadilhas fotográficas sem aprendizado profundo

Diferentemente de todos os trabalhos descritos até aqui, Wei et al. (2020) desenvolveram o Zilong, uma ferramenta para identificação de imagens vazias não baseada em aprendizado de máquina. O Zilong opera fazendo a comparação entre duas imagens: uma imagem de entrada; e uma imagem considerada imagem de referência, que não contém animais e que tenha sido capturada no mesmo local de captura da imagem de entrada. Na sequência, é realizado um cálculo de distância entre os valores de pixels das duas imagens, ou seja, é medido o quão diferente uma imagem é da outra. Caso a distância entre os pixels das imagens seja grande, então considera-se que a imagem de entrada possui animais, caso contrário, a imagem será considerada vazia.

Os autores realizaram uma comparação do Zilong com o modelo ResNet18 (TABAK et al., 2019). Os resultados mostraram que o Zilong identificou corretamente 87% das imagens de animais e 85% das imagens vazias, enquanto a ResNet18 identificou 65% e 69%, respectivamente. Apesar dos resultados serem promissores, Zilong possui dificuldade em processar imagens de animais que são muito lentos ou que não se movem, como coalas e crocodilos, visto que a diferença de pixels entre a imagem de entrada e a de referência pode ser muito baixa. Outra dificuldade surge de situações em que a vegetação está constantemente em movimentação, como por exemplo, regiões

com muito vento. Nesse caso, a distância entre os pixels das imagens pode ser alta mesmo sem a presença de animais.

3.3 Síntese do capítulo

A maioria dos estudos existentes na literatura e descritos neste capítulo utiliza abordagens agnósticas de aprendizagem profunda para filtrar imagens vazias em armadilhas fotográficas. Essas abordagens se destacam pela sua capacidade de processar grandes volumes de dados de forma eficiente, identificando padrões complexos e sutis que podem indicar a presença ou ausência de interesse biológico nas imagens.

Na Tabela 1, é apresentado um breve resumo dos modelos utilizados em cada trabalho, juntamente com os resultados obtidos e as métricas avaliadas. A análise desses resultados revela insights importantes sobre a eficácia e os limites das diferentes abordagens adotadas, fornecendo um panorama valioso para orientar futuras pesquisas nessa área.

É possível observar que esses trabalhos empregaram majoritariamente redes de convolução individuais ou conjuntos de redes de convolução. Essas arquiteturas são populares devido à sua capacidade de extrair características relevantes das imagens de forma automatizada e hierárquica, o que é fundamental para a tarefa de filtragem de imagens vazias em armadilhas fotográficas.

No entanto, o estudo conduzido por Wei et al. (2020) destaca que a comparação entre as imagens capturadas com uma imagem de referência vazia do mesmo local fornece informações cruciais relacionadas ao contexto no qual a câmera de armadilha fotográfica está instalada. Dessa forma, dificuldades relacionadas à iluminação do dia, estações do ano e muitas outras podem ser reduzidas devido às informações do local de captura contidas na imagem de referência. Essa abordagem contextualiza ainda mais a análise das imagens, aumentando a precisão do processo de filtragem.

Considerando esse contexto, neste trabalho foi empregado um modelo de aprendizagem profunda, mais especificamente uma rede Siamesa, para filtrar imagens vazias por meio da comparação das imagens capturadas com uma imagem vazia de referência.

A escolha desse modelo se deve à sua capacidade de aprender representações semânticas robustas e invariantes a variações de contexto, o que é crucial para lidar com as complexidades inerentes aos ambientes naturais.

Isso permite que o modelo leve em consideração as características específicas da vegetação local na imagem, adaptando-se de forma dinâmica às variações sazonais e ambientais. Além disso, a abordagem Siamesa possibilita uma análise comparativa mais precisa, identificando não apenas a presença ou ausência de objetos de interesse, mas também a similaridade entre as imagens capturadas e a imagem de referência vazia.

O próximo capítulo apresenta uma descrição abrangente do método proposto, detalhando os aspectos teóricos e práticos da implementação da rede Siamesa para a filtragem de imagens vazias em armadilhas fotográficas.

Tabela 1 – Resumo dos trabalhos relacionados

Autor	Modelo	Aprendizado profundo	Métrica	%
Norouzzadeh et al. (2018)	AlexNet, NiN, VGG, GoogLeNet, ResNet's	Sim	Acurácia	96.8
Willi et al. (2019)	ResNet18	Sim	Precisão	98
Yang et al. (2021a)	AlexNet	Sim	Erro de comissão	7.4
			Erro de omissão	7.3
Yang et al. (2021c)	Ensemble learning (AlexNet, ResNet e Inception)	Sim	Erro de omissão	0.7
Yang et al. (2021b)	AlexNet	Sim	Erro geral	2.69
			Erro de comissão	6.82
			Erro de omissão	6.45
Cunha et al. (2021)	MobileNetV2, EfficientNetB0, EfficientNetB3, SSDLite+MobileNetV2 e EfficientDet-D0	Sim	Precisão	82.89
Wei et al. (2020)	Zilong	Não	Precisão	87

4

ABORDAGEM PROPOSTA

Neste capítulo, apresentamos a nossa proposta para automatizar a tarefa de filtragem de imagens vazias em armadilhas fotográficas utilizando uma rede siamesa. Conforme mencionado anteriormente, rede siamesa é uma classe de arquitetura de rede neural que é treinada para medir a similaridade entre duas ou mais entradas. Na abordagem proposta neste trabalho, as entradas são duas imagens: uma capturada seguindo o processo padrão da câmera da armadilha fotográfica e outra que é uma imagem de referência do mesmo local de captura, como ilustra a Figura 12. A imagem de referência é gerada para representar a vegetação local, ou seja, o cenário de fundo. Portanto, trata-se de uma imagem sem a presença de animais que pode ser obtida periodicamente ao longo do dia. Por exemplo, a imagem de referência pode ser atualizada a cada hora para levar em consideração mudanças na iluminação e nas condições climáticas.

Ao processar essas duas entradas, a rede siamesa determina como saída a similaridade entre as imagens, indicando dessa forma a presença ou a ausência de animais na imagem capturada. Quando a imagem de referência e a imagem capturada são consideradas dissimilares pela rede siamesa, esse resultado indica presença de animal na imagem capturada. Caso sejam consideradas similares, a imagem capturada é classificada como vazia. Esse processo permite que o sistema filtre automaticamente imagens vazias e se concentre apenas nas imagens que contêm animais, aumentando a eficiência do processo de análise.

A imagem de referência é utilizada para fornecer à rede siamesa informações sobre os aspectos da vegetação que podem ser desconsiderados, destacando exclu-

sivamente o animal presente na imagem capturada pela câmera. Portanto, a seleção adequada das imagens de referência que comporão os pares de imagens é de extrema importância para o treinamento e para a correta avaliação da capacidade de generalização do modelo. Entretanto, as bases de dados disponíveis publicamente não possuem uma organização que considere a distribuição temporal dessas imagens. Assim, para possibilitar o treinamento da rede, torna-se necessário desenvolver uma estratégia de criação de pares de imagens. Essa etapa será abordada detalhadamente na próxima seção.

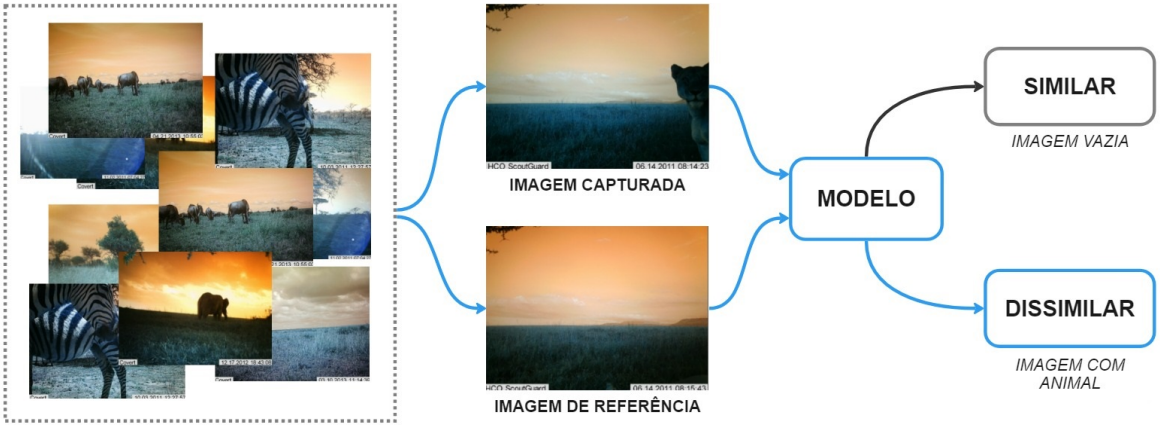


Figura 12 – Etapas da abordagem proposta. Fonte: próprio autor.

4.1 Criação de pares de imagens

Para treinar e validar a rede siamesa, é necessário montar pares de imagens. Cada par será formado por uma imagem de referência e por outra imagem (representando a imagem capturada via processo padrão da câmera), que pode conter ou não animal. Como o treinamento da rede deve simular a aplicação real do modelo, o método proposto para formar os pares de treinamento é o seguinte: cada imagem comporá um par com uma imagem de referência da mesma câmera cujo horário de captura seja o mais próximo possível do horário de obtenção da primeira imagem.

A Tabela 2 ilustra o processo de criação de pares de imagens da câmera S1_D03_R1 da base de dados Snapshot Serengeti. Por exemplo, a imagem 2 formou

par com a imagem 3, pois o tempo de captura da imagem 3 (7:22) é o mais próximo do tempo de captura da imagem 2 (7:20). É importante notar que essas duas imagens foram capturadas em dias diferentes: a imagem 2 foi obtida em 21/07/2010 e a imagem 3 em 25/07/2010. Isso ocorre porque nem sempre é possível formar pares de imagens capturadas no mesmo dia devido à natureza das bases de dados, e essa característica varia entre elas. Portanto, optou-se por criar pares com base na proximidade do tempo de captura, mesmo que as imagens tenham sido adquiridas em datas diferentes. No entanto, estabeleceu-se uma diferença máxima de 7 dias entre as datas de captura para levar em consideração os fatores ambientais que podem causar variações significativas na vegetação ao longo do tempo.

Tabela 2 – Exemplos de pares gerados com imagens da câmera S1_D03_R1 da base de dados Snapshot Serengeti.

Imagem	Câmera	Classe	Momento da captura	Par (referência)
Imagem 1	S1_D03_R1	Animal	7/21/2010 7:02	**Imagem 3**
Imagem 2	S1_D03_R1	Animal	7/21/2010 7:20	**Imagem 3**
Imagem 3	S1_D03_R1	Vazio	7/25/2010 7:22	**Imagem 2**
Imagem 4	S1_D03_R1	Vazio	7/25/2010 7:31	**Imagem 3**
Imagem 5	S1_D03_R1	Vazio	7/25/2010 9:13	**Imagem 6**
Imagem 6	S1_D03_R1	Animal	7/26/2010 9:15	**Imagem 5**
Imagem 7	S1_D03_R1	Vazio	7/26/2010 10:06	**Imagem 8**
Imagem 8	S1_D03_R1	Animal	7/26/2010 10:28	**Imagem 7**
Imagem 9	S1_D03_R1	Vazio	7/27/2010 11:45	**Imagem 10**
Imagem 10	S1_D03_R1	Animal	7/27/2010 13:01	**Imagem 9**

Diversos fatores ambientais podem gerar essas variações, como ilustrado na Figura 13, que mostra as mudanças na vegetação durante as estações do ano, e na Figura 14, que demonstra as variações na iluminação ao longo do dia. Ao criar pares de imagens levando em conta as variações nos horários do dia, buscamos treinar a rede siamesa para generalizar melhor na aplicação real. Isso se deve ao fato de que, dependendo do horário do dia, as imagens podem apresentar diferenças visuais significativas, mesmo sendo capturadas pela mesma câmera e no mesmo local. Essa diferença ocorre devido às variações na iluminação diurna, que afetam a cor dos pixels nas imagens.

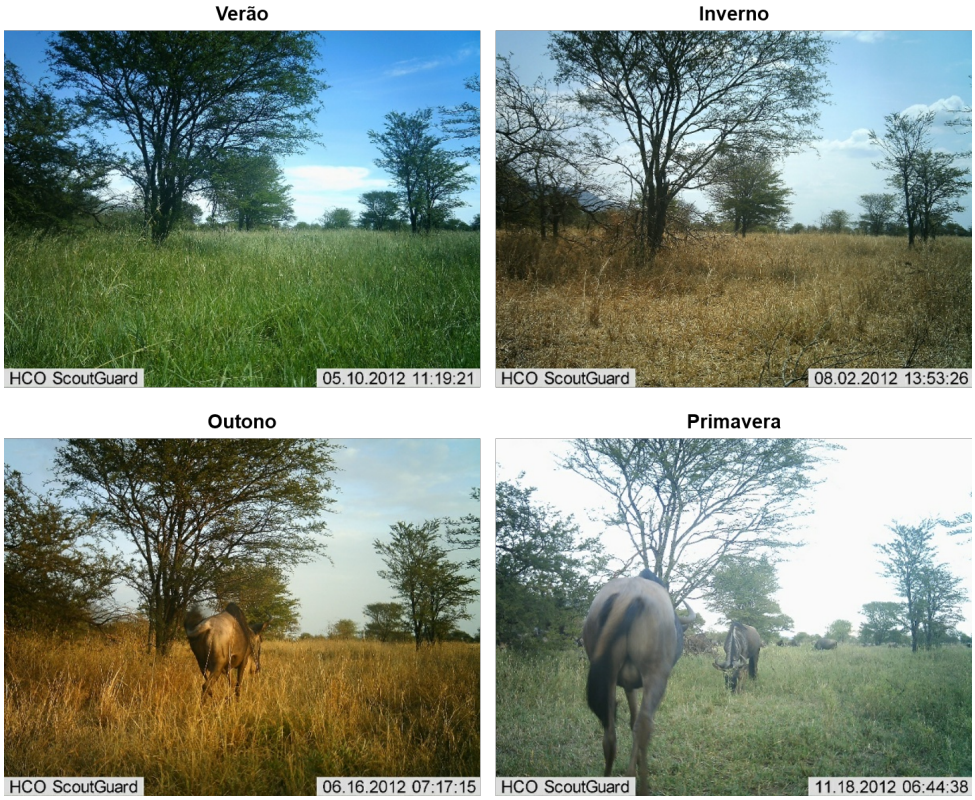


Figura 13 – Variação da vegetação do ponto de captura F02 durante as quatro estações do ano. Fonte: Base de dados Snapshot Serengeti.

Figura 14 – Variação da iluminação do ponto de captura F02 durante o dia. A primeira imagem mostra o local de manhã, a imagem do meio mostra o local à tarde, enquanto a terceira imagem mostra o local à noite. Fonte: Base de dados Snapshot Serengeti.

Uma vez criados dos pares de imagem, a próxima etapa do trabalho envolve treinamento, validação e teste da rede siamesa. Essa etapa é detalhada a seguir.

4.2 Rede Siamesa

Os pares de imagens definidos anteriormente foram utilizados para treinar três redes siamesas. Através da formação desses pares, o objetivo de cada rede siamesa é identificar as mudanças semânticas entre a imagem de referência e as imagens capturadas. Neste trabalho, as três versões de redes siamesas criadas são derivadas dos seguintes modelos de CNN:

- EfficientNetB0 (TAN; LE, 2019)
- MobileNetV2 (SANDLER et al., 2018)
- ResNet50 (HE et al., 2016a)

Os dois primeiros modelos foram selecionados por serem considerados eficientes em termos de custo computacional. Isso significa que eles podem ser executados em dispositivos com recursos limitados, como dispositivos móveis ou dispositivos com capacidade de processamento limitado, que podem ser embarcados na própria câmera da armadilha fotográfica. No caso da ResNet50, este modelo foi escolhido por ser uma arquitetura tradicional de CNN comumente utilizada como *baseline* em trabalhos de classificação de imagens (CUNHA; SANTOS; COLONNA, 2023).

A Figura 15 ilustra a arquitetura geral das redes siamesas utilizadas neste trabalho. Conforme mencionado anteriormente, nos três modelos de redes siamesas criados, as componentes CNN são redes idênticas.

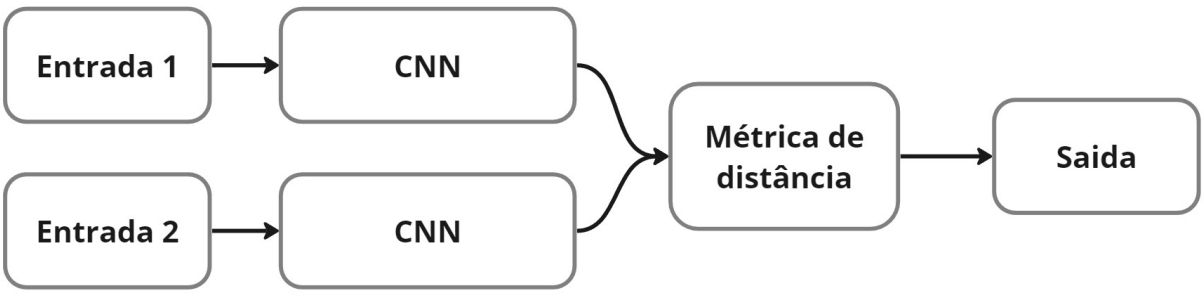


Figura 15 – Arquitetura geral da rede siamesa. Fonte: próprio autor.

Além das três versões de redes siamesas criadas, cada modelo de CNN foi utilizado para classificar as imagens de forma convencional, ou seja, como um classificador individual. A Figura 16 apresenta uma representação visual do processo realizado por cada rede siamesa criada ao processar um par de imagens, com o objetivo de calcular a similaridade entre as suas respectivas entradas. Além disso, a imagem à direita na Figura 16 mostra o funcionamento de rede CNN convencional, que processa apenas uma imagem como entrada.

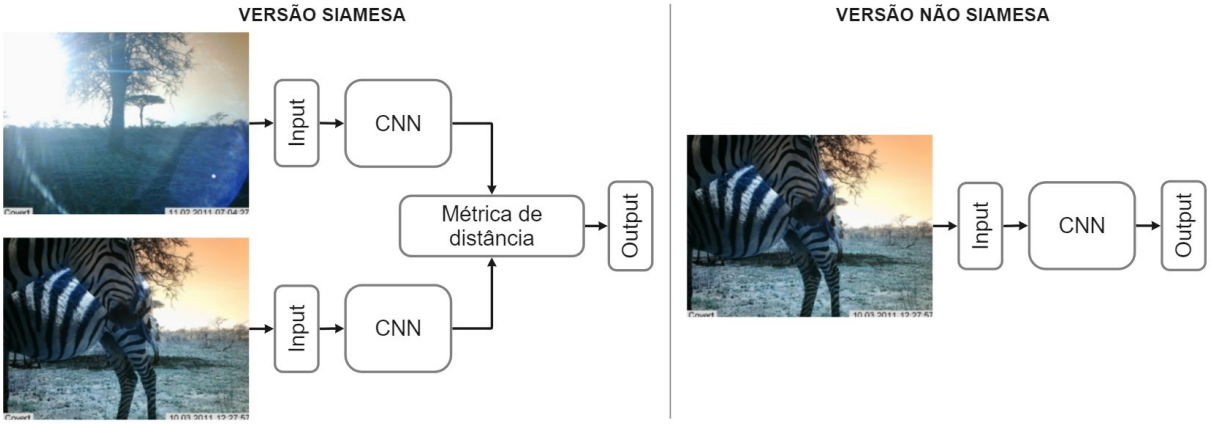


Figura 16 – Processo de classificação de imagens de armadilha fotográfica utilizando uma rede siamesa (esquerda) e uma rede CNN não siamesa (direita). Fonte: próprio autor.

4.2.1 Função de perda

A função de perda escolhida para o treinamento das redes siamesas foi a perda constrastiva (*contrastive loss*). Essa função é comumente usada em redes siamesas para aprendizado de representação, especialmente em tarefas de reconhecimento de padrões, como reconhecimento facial, verificação e alinhamento de texto (NECULOIU; VERSTEEGH; ROTARU, 2016). O objetivo do uso da perda constrastiva é aprender representações que mantenham distâncias pequenas entre exemplos similares e distâncias grandes entre exemplos diferentes no espaço de representação.

A fórmula geral para a perda contrastiva é mostrada na equação abaixo (Equação 4.1):

$$L(W,(Y,X_1,X_2)^i) = (1-Y)Ls(Dw^i) + YLd(Dw^i) \tag{4.1}$$

O termo Y especifica se os dois pontos de dados fornecidos (X_1 e X_2) são semelhantes (Y = 0) ou diferentes (Y = 1). O termo Ls representa a função de perda, que deve ser aplicada à saída se as amostras fornecidas forem semelhantes, o termo Ld - uma função de perda a ser aplicada quando os pontos de dados fornecidos forem diferentes. O termo Dw entre parênteses é a semelhança entre 2 pontos de dados transformados, dados por Chopra, Hadsell e LeCun (2005) da seguinte forma (Equação 4.2):

$$Dw(X_1,X_2) = \|Gw(X_1) - Gw(X_2)\|_2 \tag{4.2}$$

O termo G nesta fórmula representa a própria função de mapeamento, ou seja, uma rede neural. Essa é uma função de distância euclidiana regular (calculada entre as saídas da rede neural), a mesma usada em (CHOPRA; HADSELL; LECUN, 2005).

4.2.2 Configurações do treinamento do modelo

A abordagem proposta utilizando rede siamesa requer ajustes de hiperparâmetros, como é comum em modelos de redes neurais. As entradas dos três modelos utilizados para compor as redes siamesas possuem dimensões [256, 256, 3], representando imagens coloridas. Portanto, as imagens das diferentes bases de dados utilizadas nos experimentos foram redimensionadas antes de serem alimentadas na rede. Além disso, outros hiperparâmetros que precisam ser configurados são o otimizador, a taxa de aprendizagem, o número de épocas de treinamento e o tamanho do lote. Para os experimentos realizados, descritos no próximo capítulo, os valores desses hiperparâmetros são os seguintes:

- Otimizador: AdamW
- Taxa de aprendizagem: 0,001
- Número de épocas: 20
- Tamanho de mini-lote: 8

A definição dos hiperparâmetros foi conduzida por meio da exploração de diversas combinações de valores, buscando identificar a configuração que mais otimiza o desempenho da rede siamesa.

4.2.3 Aumento de dados

Para aumentar a diversidade das instâncias de treinamento e melhorar o desempenho do modelo, foi adotada uma abordagem de aumento de dados. Após o carregamento dos pares de imagens na memória, foi aplicada uma transformação de espelhamento aleatório horizontal nas imagens. Além disso, cada imagem foi submetida a um recorte retangular aleatório, com relação de aspecto e área amostrada entre [3/4, 4/3] e [65%, 100%], respectivamente. Essas técnicas de aumento de dados introduzem variações adicionais nos exemplos de treinamento, aumentando a robustez e a capacidade do modelo de lidar com diferentes cenários e condições.

4.3 Síntese do capítulo

A proposta deste trabalho é utilizar uma rede neural siamesa para identificar as diferenças semânticas entre as imagens capturadas pelo sensor da câmera e as imagens de referência, que contêm as características da vegetação local e são fotografadas periodicamente ao longo do dia. Dessa forma, o modelo pode ser empregado para descartar imagens sem diferenças semânticas, ou seja, aquelas em que a imagem capturada é semelhante à imagem vazia capturada periodicamente. Ao descartar essas imagens, a câmera economiza espaço na memória, prolongando seu tempo de captura. Além disso, em pontos de captura com vegetações visualmente complexas, abordagens tradicionais de aprendizado profundo têm dificuldade em identificar os animais (SINGH et al., 2020). Espera-se que a abordagem proposta, utilizando rede neural siamesa, permita ao modelo identificar os animais em imagens com fundos complexos, uma vez que uma imagem do par sempre apresenta as características da vegetação local. Os experimentos realizados utilizando a abordagem proposta e os resultados obtidos são descritos no

próximo capítulo.

5

EXPERIMENTOS E RESULTADOS

Neste capítulo são apresentados os resultados dos experimentos realizados com a abordagem descrita no capítulo anterior. Antes, porém, a Seção 5.1 descreve as bases de dados empregadas nos experimentos. Na sequência, a Seção 5.2 apresenta detalhes do protocolo experimental empregado. Por fim, na Seção 5.3 são apresentados e discutidos os resultados dos experimentos.

5.1 Bases de dados

Os experimentos foram realizados utilizando três diferentes bases de dados públicas de imagens de armadilhas fotográficas: Snapshot Serengeti, Caltech e Wildlife Conservation Society (WCS), as quais são descritas a seguir.

5.1.1 WCS

A base de dados WCS compreende aproximadamente 1,4 milhão de imagens de armadilhas fotográficas, mostrando a notável diversidade de cerca de 675 espécies provenientes de 12 países. Essa base se destaca como um dos conjuntos de dados de armadilhas fotográficas mais abrangentes e acessíveis ao público. Os dados para esta coleção foram fornecidos pela Wildlife Conservation Society . A Figura 17a) mostra um exemplo de imagem da base WCS. [1]

[1] para mais informações WCS: https://lila.science/datasets/wcscameratraps

5.1.2 Caltech

A base Caltech apresenta uma coleção de 243.100 imagens capturadas em 140 diferentes locais da região do sudoeste dos Estados Unidos. Essa base possui rótulos detalhados de 21 categorias de animais, incluindo espécies comumente observadas como gambá, guaxinim e coiote. Além disso, abrange aproximadamente 66.000 anotações de caixa delimitadora. É importante mencionar que cerca de 70% das imagens dessa base são rotuladas como vazias, fornecendo informações valiosas sobre a ausência de animais nas áreas monitoradas. Um exemplo de imagem da base Caltech pode ser vista na Figura 17b). [2]

5.1.3 Snapshot Serengeti

A base Snapshot Serengeti contém uma extensa compilação de aproximadamente 2,65 milhões de sequências de imagens de armadilhas fotográficas, totalizando 7,1 milhões de imagens. Essas imagens foram capturadas ao longo das onze primeiras temporadas do projeto Snapshot Serengeti, que serve como principal iniciativa da rede Snapshot Safari. Ao empregar protocolos padronizados de captura de câmeras em várias áreas protegidas na África, os membros do Snapshot Safari facilitaram as comparações entre locais para avaliar a eficácia dos programas de conservação e restauração. O Parque Nacional Serengeti, na Tanzânia, é conhecido por suas inspiradoras migrações anuais de gnus e zebras, que desempenham um papel vital na condução do ciclo dinâmico de seu ecossistema. A Figura 17c) mostra uma imagem da base Serengeti. [3]

5.2 Protocolo Experimental

Considerando o extenso volume de imagens que compõem a base de dados Serengeti (7,1 milhões de imagens), foi necessário reduzir a quantidade de imagens para tornar viável a realização dos experimentos. Portanto, um subconjunto de 309.000 pares de

[2] para mais informações sobre o Caltech: https://lila.science/datasets/caltech-camera-traps
[3] para mais informações sobre o Serengeti: https://lila.science/datasets/snapshot-serengeti

(a) (b) (c)

Figura 17 – Exemplo de imagens de cada base de dados utilizada nos experimentos: (a) WCS; (b) Caltech; (c) Serengeti. Fonte: https://lila.science/datasets

Tabela 3 – Subconjuntos de imagens usados para treinamento, validação e teste em cada base de dados

Classe	Treino	Validação	Teste	Database
vazio	105000	15000	39000	serengeti
animal	105000	15000	30000	serengeti
vazio	17500	2500	6500	caltech
animal	11766	2500	5000	caltech
vazio	17500	2500	6500	wcs
animal	17500	2499	5000	wcs

imagens dessa base foi selecionado para fins de treinamento, validação e teste dos modelos. Além disso, foi necessário também reduzir a quantidade de imagens nas bases Caltech e WCS devido à presença significativa de imagens vazias. Essa excessiva quantidade de imagens vazias resultou em desequilíbrio entre as classes de imagens nos pares correspondentes. Para atingir um equilíbrio, optou-se por utilizar 45.766 pares de imagens para a base Caltech e 51.499 pares para a base WCS.

A tabela 3 fornece uma visão geral da distribuição de imagens nos conjuntos de treino, teste e validação de cada base de dados investigada, juntamente com o número correspondente de pares de imagens por classe. O valor de classe 1 denota um par de imagens que consiste em uma imagem vazia e uma imagem contendo um animal, enquanto um valor de classe 0 indica que o par de imagens compreende apenas imagens vazias, sem nenhum animal.

É importante destacar também que diversos trabalhos na literatura mostram que a divisão aleatória de instâncias entre os conjuntos de treino e de teste deve ser evitada por ocasionar uma avaliação super otimista do modelo (BEERY; HORN; PERONA, 2018).

Isso ocorre porque geralmente não é possível alcançar elevadas taxas de classificação quando modelos treinados em um local são utilizados para classificar imagens de novos locais (CUNHA et al., 2021). É difícil alcançar altas taxas de classificação mesmo em imagens de novos pontos de captura da mesma região de captura das imagens utilizadas para treinar os modelos. Devido a esse fato, para prover uma avaliação mais acurada da real capacidade de generalização dos modelos, alguns trabalhos propõem a utilização de divisão de treinamento e teste baseada em locais de captura (SCHNEIDER et al., 2020). Nesse caso, as imagens utilizadas para teste são provenientes de locais de captura diferentes dos locais de captura dos dados de treinamento. Portanto, para mitigar a tendência do modelo de aprender apenas a partir de locais específicos, neste trabalho cada partição de dados contém imagens de locais distintos. Em outras palavras, as imagens de treinamento são provenientes de câmeras instaladas em locais diferentes dos dados das partições de validação e de teste. O mesmo tratamento foi feito entre as partições de validação e de teste.

Outro ponto importante dos nossos experimentos é que os modelos de CNN pré-treinados empregados foram submetidos a ajuste fino na base de dados Serengeti por meio do seguinte processo. Primeiramente, cada modelo carregou pesos do projeto ImageNet e foi treinado na base Serengeti. Durante esse treinamento, os pesos das redes *backbones* foram descongelados para refino do processo de extração de características. Isso foi feito tanto para as versões siamesas quanto para as versões não siamesas. Na sequência, aproveitando os pesos ajustados na base Serengeti, os modelos foram treinados usando as bases de dados Caltech e WCS, sem descongelar os pesos das redes *backbones*. Essa abordagem possibilitou a utilização dos atributos mais específicos da tarefa de filtragem de imagens vazias obtidos durante o ajuste fino na base de dados Serengeti para aprimorar a transferência de aprendizado nas bases Caltech e WCS.

Vale destacar que nós também conduzimos experimentos nos quais os modelos treinados nas bases Caltech e WCS não utilizaram os pesos ajustados na base Serengeti. No entanto, nós desconsideramos esses resultados devido ao fato de a precisão dos modelos sem o ajuste fino ter sido pior do que os resultados com o ajuste fino. Uma possível razão para esse comportamento pode ser a diferença significativa entre o

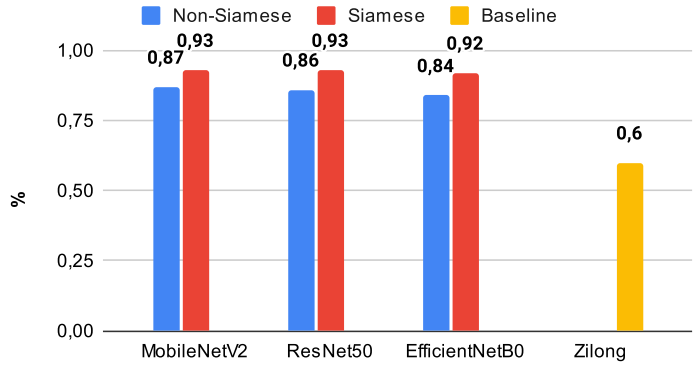
número de imagens da base Serengeti e o número de imagens das demais bases de dados, as quais são bem menores do que a base Serengeti.

5.3 Resultados

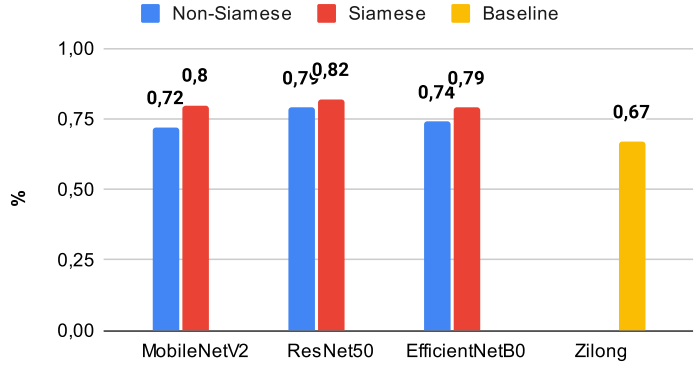
Em nossos experimentos nós comparamos o desempenho das três redes siamesas criadas com os modelos EfficientNetB0, MobileNetV2 e ResNet50 como *backbones*. Além disso, nós também testamos as versões não siamesa de cada modelo. Por fim, também foi avaliado o modelo Zilong (WEI et al., 2020) (apresentado no Capítulo 2), que não usa aprendizado profundo, mas pode processar duas imagens simultaneamente como entrada para determinar se existe ou não animal nas imagens. Neste projeto de mestrado, o Zilong foi usado como um *baseline* importante devido ao seu funcionamento relativamente parecido com a abordagem usada nas redes siamesas criadas. Para fins de comparação, os pares de imagens usados no teste do Zilong foram exatamente os mesmos usados nos outros modelos.

A Figura 18 mostra a acurácia de cada modelo avaliado na partição de teste de todas as bases de dados utilizadas nos experimentos. É evidente que os modelos siamesas superaram os modelos não siamesas em termos de acurácia. Essa diferença significativa destaca a superioridade dos modelos siamesas, que consistentemente alcançaram taxas de sucesso mais altas em comparação com suas versões não siamesas e com o Zilong. É importante destacar o baixo desempenho do método Zilong, que obteve os menores valores de acurácia dentre todos os modelos testados em todas as bases de dados investigadas. Esse baixo desempenho do Zilong confirma suas limitações apontadas na literatura, tais como: dificuldade em detectar animais que são muito lentos ou que não se movem, e dificuldade com imagens de locais cuja vegetação está constantemente em movimentação.

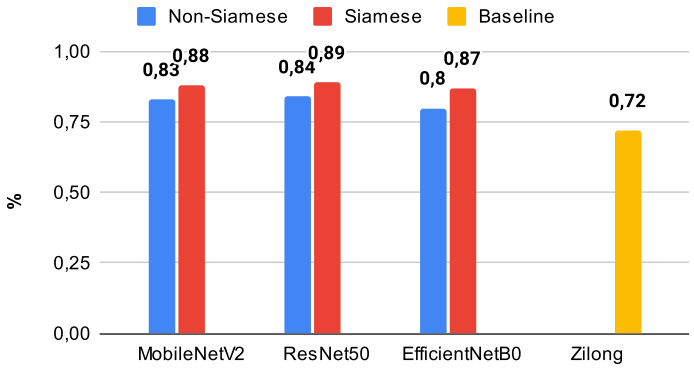
As tabelas 4, 5 e 6 apresentam métricas detalhadas para cada modelo nas bases de dados Serengeti, Caltech e WCS, respectivamente. Os valores em negrito destacam as maiores taxas obtidas pelos modelos. Ao analisar os resultados obtidos na base de dados Serengeti, observa-se que os modelos siamesas não demonstraram superioridade



Resultados da base de dados Serengeti.



Resultados da base de dados Caltech.



Resultados da base de dados do WCS.

Figura 18 – Acurácia dos modelos em todas as partições de teste de todas as bases de dados.

em todas as métricas quando comparadas às suas versões não siamesas. Por exemplo, no caso do MobileNetV2-Siamesa, a precisão para a classe "vazia"foi de 90%, enquanto a versão não siamesa alcançou uma precisão de 95% para a mesma classe. De maneira semelhante, a ResNet50-Siamesa obteve precisão de 93% para a classe "vazia", enquanto a versão não siamesa alcançou uma precisão maior (96%). No entanto, ao examinarmos os resultados alcançados nas bases de dados Caltech e WCS, fica evidente que os modelos

Tabela 4 – Resultados obtidos usando a base de dados Serengeti

Modelo	Classe	Precisão	Revocação	F1-Score
MobileNetV2-Siamesa	animal	**0.97**	0.86	**0.91**
MobileNetV2	animal	0.79	**0.94**	0.86
MobileNetV2-Siamesa	vazio	0.90	**0.98**	**0.94**
MobileNetV2	vazio	**0.95**	0.81	0.87
ResNet50-Siamesa	animal	**0.93**	0.90	**0.92**
ResNet50	animal	0.77	**0.96**	0.85
ResNet50-Siamesa	vazio	0.93	**0.95**	**0.94**
ResNet50	vazio	**0.96**	0.78	0.86
EfficientNetB0-Siamesa	animal	**0.91**	0.90	**0.90**
EfficientNetB0	animal	0.77	0.90	0.83
Zilong	animal	0.53	0.70	0.61
Zilong	vazio	0.70	0.53	0.60

siamesas superaram consistentemente todas as variações não siamesas, demonstrando sua superioridade geral.

Em aplicações práticas de armadilhas fotográficas, a captura do máximo de imagens de animais é fundamental. Por isso, a principal métrica para avaliar o desempenho dos modelos no mundo real é a precisão na classificação das classes animais. Em todas as bases de dados analisadas, os modelos de rede siamesa superaram suas contrapartes não siamesas em termos de precisão na classificação animal. Na base Serengeti, a MobileNetV2 obteve a melhor precisão (97%), seguida pela ResNet50 (86%) na base Caltech e pela EfficientNetB0 (93%) na base WCS.

O método Zilong apresentou desempenho inferior aos modelos siamesas e não siamesas, especialmente na precisão de animais da base Serengeti, atingindo apenas 53%. Um dos possíveis motivos pode estar relacionado à complexidade das imagens com animais. Dependendo da qualidade e nitidez dos animais na imagem, bem como da variação de luz, métodos que não são baseados em aprendizado profundo podem ter dificuldade em classificar as imagens. Isso ocorre porque esses métodos não conseguem abstrair alguns padrões ou representações de objetos em imagens. Em particular, armadilhas fotográficas podem gerar imagens com *backgrounds* extremamente complexos, dificultando a classificação dos animais presentes.

Tabela 5 – Resultados obtidos usando a base de dados Caltech

Modelo	Classe	Precisão	Revocação	F1-Score
MobileNetV2-Siamesa	animal	**0.84**	**0.66**	**0.74**
MobileNetV2	animal	0.73	0.56	0.63
MobileNetV2-Siamesa	vazio	**0.78**	**0.90**	**0.83**
MobileNetV2	vazio	0.71	0.84	0.77
ResNet50-Siamesa	animal	**0.86**	**0.69**	**0.77**
ResNet50	animal	0.80	0.67	0.73
ResNet50-Siamesa	vazio	**0.79**	**0.91**	**0.85**
ResNet50	vazio	0.78	0.87	0.82
EfficientNetB0-Siamesa	animal	**0.85**	**0.62**	**0.72**
EfficientNetB0	animal	0.74	0.61	0.67
EfficientNetB0-Siamesa	vazio	**0.76**	**0.91**	**0.83**
EfficientNetB0	vazio	0.74	0.83	0.78
Zilong	animal	0.64	0.56	0.60
Zilong	vazio	0.69	0.75	0.72

Tabela 6 – Resultados obtidos usando a base de dados WCS

Modelo	Classe	Precisão	Revocação	F1-Score
MobileNetV2-Siamesa	animal	**0.90**	**0.82**	**0.86**
MobileNetV2	animal	0.80	0.81	0.80
MobileNetV2-Siamesa	vazio	**0.87**	**0.93**	**0.90**
MobileNetV2	vazio	0.85	0.84	0.85
ResNet50-Siamesa	animal	**0.89**	**0.85**	**0.87**
ResNet50	animal	0.83	0.79	0.81
ResNet50-Siamesa	vazio	**0.89**	**0.92**	**0.90**
ResNet50	vazio	0.85	0.88	0.86
EfficientNetB0-Siamesa	animal	**0.93**	0.75	**0.83**
EfficientNetB0	animal	0.77	**0.79**	0.78
EfficientNetB0-Siamesa	vazio	**0.83**	**0.96**	**0.89**
EfficientNetB0	vazio	0.83	0.82	0.83
Zilong	animal	0.71	0.61	0.66
Zilong	vazio	0.73	0.81	0.77

Tabela 7 – Resultados dos modelos siamesas

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**	**Base**
MobileNetV2-Siamesa	**0.93**	**0.93**	0.92	0.92	Serengeti
ResNet50-Siamesa	**0.93**	**0.93**	**0.92**	**0.93**	Serengeti
EfficientNetB0-Siamesa	0.92	0.91	0.91	0.91	Serengeti
MobileNetV2-Siamesa	0.80	0.81	0.78	0.78	Caltech
ResNet50-Siamesa	**0.82**	**0.82**	**0.80**	**0.81**	Caltech
EfficientNetB0-Siamesa	0.79	0.80	0.76	0.77	Caltech
MobileNetV2-Siamesa	0.88	0.88	**0.87**	0.88	WCS
ResNet50-Siamesa	**0.89**	**0.89**	**0.88**	**0.88**	WCS
EfficientNetB0-Siamesa	0.87	0.88	0.85	0.86	WCS

A Tabela 7 resume os resultados dos modelos siamesas em todas as bases de dados. É importante notar que a ResNet50 foi o modelo que alcançou as maiores taxas em todas as bases, provavelmente devido à sua arquitetura mais complexa. Porém, os resultados obtidos por todos os modelos siamesas foram relativamente próximos.

5.4 Síntese do capítulo

Este capítulo apresentou detalhes sobre a criação dos modelos de redes siamesas e o seu treinamento realizado em três diferentes bases de dados de imagens de armadilha fotográfica. Para garantir a diversidade e uma avaliação mais acurada da capacidade de generalização dos modelos, foi assegurado que as imagens de treinamento não fossem provenientes do mesmo local de captura das imagens de validação e de teste. Além disso, os pesos obtidos via ajuste fino durante o treinamento da base de dados Serengeti foram aproveitados para melhorar o desempenho dos modelos nas bases Caltech e WCS. Ao analisar os resultados, constatou-se que os modelos siamesas apresentaram consistentemente um desempenho superior em comparação às versões não siamesas nas bases de dados Caltech e WCS, embora tenham ocorrido algumas variações nos resultados da base de dados Serengeti. Em relação ao Zilong, entretanto, tanto os modelos siamesas quanto não siamesas superaram o Zilong em todos os experimentos realizados.

6

Conclusão

Esta dissertação descreve um método de classificação baseado em rede neural convolucional siamesa (CNN) proposto para filtrar imagens vazias em dados de armadilhas fotográficas. A rede siamesa proposta compara imagens capturadas por câmeras de armadilhas fotográficas com imagens de referência do mesmo local de captura para determinar se a imagem capturada contém um animal ou não. Essa abordagem aproveita as características específicas da vegetação local presente nas imagens de referência para permitir ao modelo diferenciar imagens contendo animais de imagens que não apresentam animais.

Este trabalho também descreveu o processo de criação dos pares de imagens utilizados para treinar a rede siamesa. Os pares foram criados levando-se em consideração a distribuição temporal entre as imagens e as variações ambientais. Também foram aplicadas técnicas de aumento de dados para incrementar a diversidade de instâncias de treinamento e melhorar a robustez do modelo criado.

Os experimentos foram conduzidos usando três bases de dados públicas: Snapshot Serengeti, Caltech e WCS. Além disso, três modelos de CNN foram usados para gerar três redes siamesas. Os modelos de CNN empregados foram: MobileNetV2, ResNet50 e EfficientNetB0. Todas as três redes siamesas criadas obtiveram resultados competitivos nas diferentes bases de dados investigadas. O método proposto foi comparado ao uso das CNNs de forma individual e a um *baseline* que também faz a comparação entre uma imagem de referência e as imagens capturadas, porém, de forma não supervisionada. Os resultados mostraram um desempenho superior da abordagem de rede siamesa

quando comparada aos demais métodos testados na classificação precisa de imagens vazias e de animais.

O método proposto fornece uma solução prática para enfrentar o desafio da filtragem de imagens vazias em dados de armadilhas fotográficas. Ao filtrar efetivamente as imagens vazias, os pesquisadores e conservacionistas podem concentrar seus esforços na análise do comportamento animal e na estimativa do tamanho da população de espécies específicas, levando a um monitoramento mais preciso e eficiente da vida selvagem e à conservação ecológica. Durante o desenvolvimento deste trabalho, foi possível publicar um artigo com parte dos resultados no SIBGRAPI (Conference on Graphics, Patterns and Images) (ALENCAR; CUNHA; SANTOS, 2023), tendo recebido o prêmio de "Best Paper"na trilha principal.

6.1 Limitações

A eficácia da abordagem proposta está diretamente relacionada à disponibilidade de imagens de referência do mesmo local onde as imagens foram capturadas pelas armadilhas fotográficas. Caso a obtenção dessas imagens de referência seja difícil ou inviável, a aplicabilidade prática da abordagem pode ser limitada. Além disso, é importante reconhecer que o trabalho foi avaliado em ambientes específicos, como as bases de dados Snapshot Serengeti, Caltech e WCS. Embora os resultados tenham sido promissores, a generalização para outros ambientes com características diferentes pode ser um desafio. Por exemplo, ambientes com vegetação ainda mais densa e complexa podem apresentar desafios adicionais que não foram abordados no trabalho. Contudo, é importante ressaltar que as bases de dados Caltech e WCS possuem imagens extremamente complexas devido à densidade de *habitats*, o que pode contribuir para a abrangência da abordagem (O'BRIEN; KINNAIRD; WIBISONO, 2003).

Outra possível limitação a ser considerada é a qualidade e precisão das imagens de referência utilizadas. O desempenho da rede siamesa pode ser afetado negativamente caso as imagens de referência contenham artefatos ou imprecisões que dificultem a correta comparação com as imagens capturadas pelas armadilhas fotográficas. Portanto,

a seleção cuidadosa e a obtenção de imagens de referência de alta qualidade são aspectos cruciais a serem considerados na aplicação prática da abordagem.

Essas limitações ressaltam a importância de futuras investigações que busquem melhorar a flexibilidade e adaptabilidade da abordagem, bem como aprimorar a coleta e a qualidade das imagens de referência. Ao enfrentar esses desafios, será possível fortalecer a eficácia e a utilidade da abordagem na tarefa crucial de filtragem de imagens vazias em projetos de monitoramento e conservação da vida selvagem.

6.2 Trabalhos futuros

Este trabalho abriu diversas possibilidades de extensão e aprimoramento da abordagem proposta para a filtragem de imagens vazias em dados de armadilhas fotográficas. Algumas das principais direções para trabalhos futuros são:

- Exploração de outras arquiteturas de redes neurais: Investigar o desempenho de outras arquiteturas de redes neurais convolucionais siamesas, além das utilizadas neste trabalho (MobileNetV2, ResNet50 e EfficientNetB0). Diferentes arquiteturas podem oferecer vantagens específicas para a tarefa de filtragem de imagens vazias em ambientes com fundos complexos.

- Integração de informações adicionais: Investigar a incorporação de informações adicionais, como dados meteorológicos, informações sobre a biodiversidade local ou características específicas dos animais alvo, para melhorar a precisão da filtragem de imagens vazias.

- Aprendizado semi - supervisionado: Explorar técnicas de aprendizado semi - supervisionado para aproveitar melhor os dados não rotulados disponíveis nos conjuntos de dados de armadilhas fotográficas, ampliando assim a capacidade do modelo de se adaptar a novos cenários.

- Avaliação em ambientes específicos: Realizar experimentos em ambientes específicos, como florestas tropicais, savanas ou áreas de alta densidade de vegetação,

para avaliar o desempenho da abordagem em cenários desafiadores e garantir sua robustez em diferentes contextos.

- Combinação com outras técnicas de visão computacional: Investigar a combinação da abordagem baseada em redes siamesas com outras técnicas de visão computacional, como segmentação de imagens ou detecção de objetos, para aprimorar a filtragem de imagens vazias.

- A aplicação de técnicas de compactação, quantização e compressão de modelos visando aprimorar a eficiência computacional para que as redes possam ser incorporadas em câmeras de armadilhas fotográficas, reduzindo os custos computacionais associados.

REFERÊNCIAS

ABIODUN, O. I. et al. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, Elsevier, v. 4, n. 11, p. e00938, 2018. 24

ALENCAR, L.; CUNHA, F.; SANTOS, E. M. dos. A context-aware approach for filtering empty images in camera trap data using siamese network. In: IEEE. *2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.], 2023. p. 85–90. 18, 62

AUER, D. et al. Minimizing the annotation effort for detecting wildlife in camera trap images with active learning. *INFORMATIK 2021*, Gesellschaft für Informatik, Bonn, 2021. 17

BEERY, S.; HORN, G. V.; PERONA, P. Recognition in terra incognita. In: *Proceedings of the European conference on computer vision (ECCV)*. [S.l.: s.n.], 2018. p. 456–473. 13, 14, 15, 37, 54

BERTINETTO, L. et al. Fully-convolutional siamese networks for object tracking. In: SPRINGER. *European conference on computer vision*. [S.l.], 2016. p. 850–865. 16, 31

CAO, S.; NEVATIA, R. Exploring deep learning based solutions in fine grained activity recognition in the wild. In: IEEE. *2016 23rd International Conference on Pattern Recognition (ICPR)*. [S.l.], 2016. p. 384–389. 26

CHEN, X.; YUILLE, A. L. Articulated pose estimation by a graphical model with image dependent pairwise relations. *Advances in neural information processing systems*, v. 27, 2014. 26

CHEN, Z. et al. Siamese box adaptive network for visual tracking. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 6668–6677. 16, 31

CHOPRA, S.; HADSELL, R.; LECUN, Y. Learning a similarity metric discriminatively, with application to face verification. In: IEEE. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. [S.l.], 2005. v. 1, p. 539–546. 49

CUNHA, F. et al. Filtering empty camera trap images in embedded systems. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2021. p. 2438–2446. 14, 21, 22, 39, 40, 42, 55

CUNHA, F.; SANTOS, E. M. dos; COLONNA, J. G. Bag of tricks for long-tail visual recognition of animal species in camera-trap images. *Ecological Informatics*, v. 76, p. 102060, 2023. 47

DEY, S. et al. Signet: Convolutional siamese network for writer independent offline signature verification. *arXiv preprint arXiv:1707.02131*, 2017. 16

DHILLON, A.; VERMA, G. K. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, Springer, v. 9, n. 2, p. 85–112, 2020. 24

DIBA, A. et al. Weakly supervised cascaded convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 914–922. 26

DOULAMIS, N. Adaptable deep learning structures for object labeling/tracking under dynamic visual environments. *Multimedia Tools and Applications*, Springer, v. 77, n. 8, p. 9651–9689, 2018. 26

DOULAMIS, N.; VOULODIMOS, A. Fast-mdl: Fast adaptive supervised training of multi-layered deep learning models for consistent object tracking and classification. In: IEEE. *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*. [S.l.], 2016. p. 318–323. 26

ELIAS, A. R. et al. Where's the bear?-automating wildlife image processing using iot and edge cloud systems. In: IEEE. *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*. [S.l.], 2017. p. 247–258. 14

FANG, S. et al. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 19, p. 1–5, 2021. 32

FREY, S. et al. Investigating animal activity patterns and temporal niche partitioning using camera-trap data: Challenges and opportunities. *Remote Sensing in Ecology and Conservation*, Wiley Online Library, v. 3, n. 3, p. 123–132, 2017. 13

GOLHANI, K. et al. A review of neural networks in plant disease detection using hyperspectral data. *Information Processing in Agriculture*, Elsevier, v. 5, n. 3, p. 354–371, 2018. 24

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016. 24, 25, 26

HE, A. et al. A twofold siamese network for real-time object tracking. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 4834–4843. 30

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. 28, 47

HE, Z. et al. Visual informatics tools for supporting large-scale collaborative wildlife monitoring with citizen scientists. *IEEE Circuits and Systems Magazine*, IEEE, v. 16, n. 1, p. 73–86, 2016. 13

LIN, L. et al. A deep structured model with radius–margin bound for 3d human activity recognition. *International Journal of Computer Vision*, Springer, v. 118, n. 2, p. 256–273, 2016. 26

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 3431–3440. 26

NECULOIU, P.; VERSTEEGH, M.; ROTARU, M. Learning text similarity with siamese recurrent networks. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*. [S.l.: s.n.], 2016. p. 148–157. 48

NOH, H.; HONG, S.; HAN, B. Learning deconvolution network for semantic segmentation. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 1520–1528. 26

NOROUZZADEH, M. S. et al. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 115, n. 25, p. E5716–E5725, 2018. 13, 14, 15, 20, 21, 22, 37, 42

O'BRIEN, T. G.; KINNAIRD, M. F.; WIBISONO, H. T. Crouching tigers, hidden prey: Sumatran tiger and prey populations in a tropical forest landscape. In: CAMBRIDGE UNIVERSITY PRESS. *Animal Conservation Forum*. [S.l.], 2003. v. 6, n. 2, p. 131–139. 62

OUYANG, W. et al. Deepid-net: Object detection with deformable part based convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 39, n. 7, p. 1320–1334, 2016. 26

PENG, H. et al. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, Elsevier, v. 68, p. 101871, 2021. 25

REZAEI, S.; LIU, X. Deep learning for encrypted traffic classification: An overview. *IEEE communications magazine*, IEEE, v. 57, n. 5, p. 76–81, 2019. 25

RICH, L. N. et al. Artificial water catchments influence wildlife distribution in the mojave desert. *The Journal of Wildlife Management*, Wiley Online Library, v. 83, n. 4, p. 855–865, 2019. 13

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. *International Conference on Medical image computing and computer-assisted intervention*. [S.l.], 2015. p. 234–241. 32

SANDLER, M. et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 4510–4520. 27, 47

SCHNEIDER, S. et al. Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and evolution*, Wiley Online Library, v. 10, n. 7, p. 3503–3517, 2020. 37, 55

SCHNEIDER, S. et al. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, Wiley Online Library, v. 10, n. 4, p. 461–470, 2019. 13

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 29

SINGH, P. et al. Animal localization in camera-trap images with complex backgrounds. In: IEEE. *2020 IEEE southwest symposium on image analysis and interpretation (SSIAI)*. [S.l.], 2020. p. 66–69. 8, 15, 16, 50

SONG, L. et al. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2019. p. 773–782. 31

SWANSON, A. et al. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, Nature Publishing Group, v. 2, n. 1, p. 1–14, 2015. 8, 14, 21, 22, 23

TABAK, M. A. et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, Wiley Online Library, v. 10, n. 4, p. 585–590, 2019. 13, 14, 15, 20, 21, 22, 40

TAIGMAN, Y. et al. Deepface: Closing the gap to human-level performance in face verification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2014. p. 1701–1708. 16, 31

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. *International conference on machine learning*. [S.l.], 2019. p. 6105–6114. 29, 47

TEALAB, A. Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, Elsevier, v. 3, n. 2, p. 334–340, 2018. 24

TOSHEV, A.; SZEGEDY, C. Deeppose: Human pose estimation via deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2014. p. 1653–1660. 26

WEI, W. et al. Zilong: A tool to identify empty images in camera-trap data. *Ecological Informatics*, Elsevier, v. 55, p. 101021, 2020. 14, 21, 22, 32, 40, 41, 42, 56

WILLI, M. et al. Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, Wiley Online Library, v. 10, n. 1, p. 80–91, 2019. 13, 14, 15, 21, 22, 37, 38, 42

YANG, D.-Q. et al. A systematic study of the class imbalance problem: Automatically identifying empty camera trap images using convolutional neural networks. *Ecological Informatics*, Elsevier, v. 64, p. 101350, 2021. 13, 38, 42

YANG, D.-Q. et al. An adaptive automatic approach to filtering empty images from camera traps using a deep learning model. *Wildlife Society Bulletin*, Wiley Online Library, v. 45, n. 2, p. 230–236, 2021. 21, 22, 39, 42

YANG, D.-Q. et al. An automatic method for removing empty camera trap images using ensemble learning. *Ecology and Evolution*, Wiley Online Library, v. 11, n. 12, p. 7591–7601, 2021. 21, 22, 39, 42