UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Romulo Augusto Aires Soares

Detecção de Armas de Fogo usando DETR com Múltiplas Redes Neurais Autocoordenadas

São Luís
2024

ROMULO AUGUSTO AIRES SOARES

DETECÇÃO DE ARMAS DE FOGO USANDO DETR COM MÚLTIPLAS REDES NEURAIS AUTOCOORDENADAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão, como requisito necessário para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Areolino de Almeida Neto

SÃO LUÍS - MA

2024

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Soares, Romulo Augusto Aires.
Detecção de Armas de Fogo usando DETR com Múltiplas Redes Neurais Autocoordenadas / Romulo Augusto Aires Soares. - 2024.
67 f.

Orientador(a): Areolino de Almeida Neto.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, São Luis-ma, 2024.

1. Rede Neural. 2. Múltiplas Redes Neurais. 3. Detecção de Armas. 4. Detr. 5. . I. Almeida Neto, Areolino de. II. Título.

ROMULO AUGUSTO AIRES SOARES

DETECÇÃO DE ARMAS DE FOGO USANDO DETR COM MÚLTIPLAS REDES NEURAIS AUTOCOORDENADAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão, como requisito necessário para obtenção do título de Mestre em Ciência da Computação

Dissertação aprovada em 02 de setembro de 2024.

_____

Prof. Dr. Areolino de Almeida Neto
Universidade Federal do Maranhão - UFMA
Orientador

_____

Prof. Dr. Omar Andres Carmona Cortes
Instituto Federal do Maranhão - IFMA
Examinador Interno

_____

Prof. Dr. Guilherme de Alencar Barreto
Universidade Federal do Ceará – UFC
Examinador Externo

São Luís - MA
2024

A Elis Maria Gomes Aires (In memorian),
Luis Carlos Soares (In memorian),
Maria de Jesus Soares Aires (In memorian),
José Miguel Ferreira Aires (In memorian)
e
José Miguel Ferreira Aires Junior (In memorian).

AGRADECIMENTOS

Agradeço a Deus pelas bênçãos e por ter me feito chegar até aqui.

Agradeço também à minha família pelo apoio necessário, em especial à minha mãe Luciane Aires Soares pela ajuda, amor e felicidade em todas as situações.

Aos professores passados e atuais por cada ensinamento e contribuição para meu desenvolvimento pessoal e/ou profissional.

Gostaria também de agradecer, ao meu orientador, prof. Areolino de Almeida Neto, pela paciência, motivação e sabedoria compartilhada nas reuniões que me ajudaram a caminhar esta minha fase do mestrado.

Agradeço também aos colegas do laboratório InovTec, em especial a Bianca pela ajuda e companheirismo ao longo de toda caminhada.

E por último, mas de forma alguma menos importante, agradeço aos meus amigos pelas palavras de incentivo e confiança, vocês fazem parte desta conquista.

“A felicidade às vezes é uma benção, mas geralmente é uma conquista.“ (Paulo Coelho).

RESUMO

Este trabalho apresenta uma nova estratégia que faz uso de múltiplas redes neurais em conjunto com a rede do tipo DEtection TRansformer (DETR) na detecção de armas de fogo em imagens de vigilância. O crescimento contínuo de violência por armas de fogo em todo o mundo obrigou várias agências, empresas e consumidores a implantar câmeras de vigilância de circuito fechado de TV (CFTV) na tentativa de combater essa epidemia. Porém o grande número de câmeras a serem observadas leva a sobrecarga dos operadores de CFTV, gerando cansaço e estresse, consequentemente, perda de eficiência na vigilância. A estratégia desenvolvida neste trabalho expõe uma metodologia que promove a colaboração e a autocoordenação das redes nas camadas *full connected* da DETR através da técnica de múltiplas redes neurais artificiais (MRNA) autocoordenadas, que dispensa o uso de coordenador. Essa autocoordenação consiste em fazer com que as redes sejam treinadas, uma após a outra, e as suas saídas sejam integradas sem um elemento extra chamado de coordenador. Tendo em vista melhorar a vigilância remota, a inserção de redes neurais profundas vem mostrando-se eficiente na detecção e identificação de objetos em vídeos, tendo, em diversas situações, produzido resultados mais precisos e consistentes que seres humanos. Tanto quanto se sabe, este trabalho é o primeiro a introduzir a MRNA autocoordendas em arquiteturas de detecção de objetos baseadas em transformadores para a detecção de objetos.

Palavras-Chave: Rede Neural; Múltiplas Redes Neurais; Detecção de armas; DETR.

ABSTRACT

This paper presents a new strategy that uses multiple neural networks in conjunction with the DEtection TRansformer (DETR) network to detect firearms in surveillance images. The continuous growth of gun violence around the world has forced many agencies, companies and consumers to deploy closed circuit TV (CCTV) surveillance cameras in an attempt to combat this epidemic. However, the large number of cameras to be observed leads to an overload of CCTV operators, generating fatigue and stress and, consequently, a loss of surveillance efficiency. The strategy developed in this work presents a methodology that promotes collaboration and self-coordination of networks in the full connected layers of DETR through the technique of multiple self-coordinating artificial neural networks (MANN), which does not require the use of a coordinator. This self-coordination consists of training the networks one after the other and integrating their outputs without an extra element called a coordinator. In order to improve remote surveillance, the insertion of deep neural networks has proven to be efficient in detecting and identifying objects in videos, and in various situations has produced more accurate and consistent results than human beings. To the best of our knowledge, this work is the first to introduce MANN into transform-based object detection architectures.

Keywords: Neural Network; Multiple Neural Networks; Weapon Detection; DETR.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CFTV	Circuito Fechado de Televisão, em inglês
CNN	Redes Neurais Convolucionais, em inglês
DETR	*DEtection TRansformer*
FNN	*Feed Forward Network*
FN	Falsos Negativos
FP	Falsos Positivos
F1	Score F1
IoU	Intersecção sobre União, em inglês
LSTM	Long Short Term Memory
MAP	*Mean Average Precision*
MLFPN	*Multi-Level Feature Pyramid Network*
MRNA	Múltiplas Redes Neurais Artificiais
MLP	*Multilayer Perceptron*
PLN	Processamento de Linguagem Natural, em inglês
Pr	Precisão
ResNet	*Residual Networks*
R-CNN	Rede Neural Convolucional de Região
RNNs	Redes Neurais Recorrentes
Se	Sensibilidade
TSP-RCNN	*Transformer-based Set Prediction with RCNN*
TSP-FCOS	*Transformer-based Set Prediction with FCOS*
ViT	*Vision Transformer*
VP	Verdadeiros Positivos
VN	Verdadeiros Negativos

SUMÁRIO

1 INTRODUÇÃO

Referente à violência por arma de fogo, esta se identifica como uma das maiores preocupações atuais da sociedade, despontando como alvo de inúmeros debates no âmbito público e privado. Vários estudos mostram que a arma de fogo é a principal arma usada para vários crimes, como arrombamento, roubo, furto em loja e estupro. O número de homicídios por armas de fogo tem estado, já há algum tempo, entre as principais causas de mortes precoces no Brasil, como bem demonstra o trabalho de Waiselfisz(2016).

Um indivíduo portando armas em público é um forte indicador de uma possível situação perigosa. Recentemente, houve o aumento do número de incidentes em que indivíduos ou pequenos grupos fazem uso de armas de fogo com o objetivo de ferir ou matar o maior número possível de pessoas. Essas situações criam traumas psicológicos normalmente entre crianças expostas a esses altos níveis de violência, sejam elas vítimas, perpetradores ou testemunhas, podendo experimentar efeitos psicológicos negativos a curto e longo prazo.

Entre os mais notáveis desses eventos, chamados de tiroteios em massa, estão o ataque à ilha Utoya (Noruega, 179 vítimas), o massacre de Realengo (Brasil, 13 vítimas), o massacre de Suzano (Brasil, 10 vítimas) e o ataque contra o jornal Charlie Hebdo (França, 23 vítimas). Diante desse cenário, várias agências, empresas e consumidores implantaram câmeras de vigilância de Circuito Fechado de TV (CFTV) em tentativa de combater essa situação.

No entanto, o uso de sistemas de CFTV pode apresentar alguns problemas, incluindo a distração e a fadiga visual do operador. Quando o operador é responsável por monitorar várias câmeras simultaneamente, a quantidade de informações visuais que precisa processar pode ser avassaladora. Essa sobrecarga de estímulos visuais pode levar à distração, tornando mais difícil detectar eventos relevantes ou suspeitos. Além disso, a necessidade de ficar constantemente focado nas telas pode causar fadiga visual, levando a problemas como olhos secos, dores de cabeça e dificuldade de concentração. Esses problemas podem comprometer a eficácia do sistema de CFTV.

Desta maneira, uma solução possível e inovadora seria utilizar de forma conjunta, um sistema de vigilância com um sistema artificial capaz de reconhecer armas de fogo e alertar os responsáveis em tempo curto. Assim, seria identificado o

comportamento perturbador no estágio inicial e monitoradas as atividades suspeitas cuidadosamente para que as autoridades competentes possam tomar medidas imediatas (DEEPTHI *et al*., 2019).

Estudos apontam que o aprendizado de máquina (P. M. Kumar *et al*., 2019; K.-E. Ko e K.-B, 2018) e os algoritmos avançados de processamento de imagem têm alcançado um papel dominante em vigilância inteligente e sistemas de segurança (Nikouei *et al*., 2018), (R. Xu *et al*., 2018). Além do mais, a popularidade de dispositivos inteligentes e câmeras em rede também capacitaram esse domínio.

Todo esse domínio e avanço tecnológico possibilitaram diversos métodos de solução com uso de diferentes abordagens em problemas de detecção de objetos. Dentre tais abordagens, em 2020, surgiu o DETR (DEtection TRansformer) trazendo o uso do transformador em visão computacional, realizando a detecção de objetos de ponta a ponta (Carion *et al*., 2020).

A arquitetura do DETR é simples, possuindo três elementos principais: um backbone, um transformador codificador-decodificador e uma rede Feed Forward (FFN), a qual prevê uma detecção (classe e caixa delimitadora) ou uma classe “sem objeto”. Esse modelo prevê todos os objetos de uma vez sem a necessidade de propostas de região e âncoras que normalmente são encontrados nos principais modelos de detecção, seja de um ou de dois estágios como Yolo ou Faster-RCNN, respectivamente.

Nesse contexto, este trabalho apresenta um estudo sobre uma metodologia que promove a colaboração e autocoordenação entre redes do tipo FFN encontradas na última camada do DETR através da técnica de Múltiplas Redes Neurais autocoordendas.

1.1 OBJETIVOS

1.1.1 Objetivo geral

O principal objetivo deste trabalho foi desenvolver uma técnica híbrida que envolva a rede DETR e múltiplas redes neurais autocoordenadas para detecção automática de armas de fogo em imagens.

1.1.2 Objetivos específicos

Para alcançar o objetivo geral, os seguintes objetivos específicos foram definidos:

- Avaliar a eficácia do modelo DETR com imagens de uma base pública de armas de fogo;
- Comparar a precisão da rede com os resultados encontrados na literatura e;
- Analisar a eficiência do uso de redes adicionais.

1.2 JUSTIFICATIVA / MOTIVAÇÃO

Os olhares da comunidade acadêmica voltam-se à utilização de redes neurais convolucionais para o problema de detecção de objetos, com ênfase para métodos como R-CNN (GIRSHICK *et al*., 2014), R-FCN (DAI *et al*., 2016), RetinaNet (LIN *et al*., 2018) e YOLO (REDMON *et al*., 2016). Geralmente, esses métodos têm levado a resultados que superam com boa margem os resultados obtidos com os métodos clássicos.

Apesar de as taxas de precisão desses algoritmos serem altas (especialmente para objetos minúsculos), eles ainda deixam a desejar em relação a possuir um equilíbrio entre precisão e velocidade de treinamento.

Além do mais, à medida que a busca por aprimorar a detecção de objetos avança, surgem desafios adicionais que necessitam de abordagens mais inovadoras. Um desses desafios é a detecção precisa de objetos em ambientes com baixa luminosidade ou condições adversas, onde as redes neurais podem apresentar dificuldades. Além disso, a escalabilidade desses modelos para lidar com um grande número de classes de objetos e a adaptação a cenários em constante mudança também têm sido pontos de interesse para a pesquisa nessa área.

Outroassim, a detecção precisa de objetos pequenos representa um desafio significativo para os sistemas de visão computacional, especialmente para modelos baseados em redes neurais como o DETR, também conhecido como transformador de detecção de ponta a ponta. Embora o DETR tenha demonstrado excelente desempenho na detecção de objetos em escalas maiores, sua eficácia é limitada quando se trata de objetos pequenos devido à natureza da sua arquitetura. Para mitigar essa limitação, propõe-se o uso de múltiplas redes neurais autocoordenadas, uma técnica inovadora que visa melhorar a precisão da detecção de objetos

pequenos no DETR. Este trabalho não apenas investigará a aplicabilidade e os benefícios dessa abordagem, mas também contribuirá significativamente para a evolução da detecção de objetos em escala reduzida, ampliando as capacidades do DETR em cenários práticos de visão computacional.

Portanto, este trabalho propõe explorar uma abordagem que combine as vantagens da rede DETR com técnica de combinação entre várias redes, para melhorar a robustez e a eficiência da detecção de armas de fogos em imagens. Ao enfatizar a busca por um equilíbrio ideal entre precisão e velocidade de treinamento, esta pesquisa tem como objetivo contribuir para o desenvolvimento de soluções mais eficazes e versáteis na detecção de armas em imagens, considerando tanto o desempenho em condições ideais quanto em cenários desafiadores. Através desse enfoque inovador, espera-se impulsionar o avanço da detecção de objetos e sua aplicabilidade em uma variedade de aplicações do mundo real.

1.3 TRABALHOS RELACIONADOS

Este subcapítulo apresenta trabalhos recentes e relacionados à detecção de armas utilizando a técnica múltiplas redes neurais artificiais. Os trabalhos aqui apresentados foram escolhidos após uma pesquisa por artigos relacionados, não apenas à detecção de armas de fogo, mas também de outros objetos que fazem uso de transformadores em detecção. Vale a pena mencionar que se optou, também, por trabalhos mais relevantes relacionados à detecção de armas de fogo em câmeras de vigilância que utilizam outros métodos de detecção.

Guan *et al*. (2021) apresentam uma nova arquitetura para detecção de objetos 3D, M3DETR, que combina diferentes representações de nuvens de pontos (*raw*, *voxels*, *bird-ey* e *view*) com diferentes escalas de recursos baseadas em pirâmides de recursos multiescala. O M3DETR é a primeira abordagem que unifica várias representações de nuvens de pontos, escalas de recursos, bem como modela relacionamentos mútuos entre nuvens de pontos simultaneamente usando transformadores.

No trabalho de He *et al*. (2021), foi apresentado o TransVOD, um modelo de detecção de objetos de vídeo de ponta a ponta baseado em uma arquitetura de Transformador espaço-temporal. Esse modelo demonstrou desempenho de

resultados comparáveis no benchmark do ImageNet VID. Pouco tempo depois, Zhou *et al*. (2022), fizeram melhorias na versão desenvolvida por He *et al*. (2021) e apresentaram duas versões aprimoradas do TransVOD, incluindo TransVOD++ e TransVOD Lite, atingindo resultados recordes de última geração em termos de precisão no ImageNet e também alcançando a melhor relação de velocidade e precisão.

No trabalho de Xu *et al*. (2021), foi apresentado um framework baseado em *transformer* para estimativa de textura humana 3D a partir de uma única imagem. O *transformer* proposto é capaz de explorar efetivamente as informações globais da imagem de entrada, superando as limitações de métodos existentes que são baseados exclusivamente em redes neurais. Os resultados demonstram a eficácia da proposta contra abordagens de estimativa de textura humana 3D de última geração, tanto quantitativa quanto qualitativamente.

Sun *et al*. (2020) propuseram duas soluções para melhorar o desempenho do DETR. Uma solução chama-se *Transformer-based Set Prediction with FCOS* (TSP-FCOS, Previsão de conjuntos baseado em transformador com FCOS, em inglês). A segunda solução chama-se *Transformer-based Set Prediction with R-CNN* (TSP-RCNN, Previsão de conjuntos baseado em transformador com *R-CNN,* em inglês). Os resultados experimentais mostram que os métodos propostos não apenas convergem muito mais rápido que o DETR original, mas também superam significativamente o DETR e outras linhas de base em termos de precisão de detecção.

Misra *et al*. (2021) propuseram o 3DETR, um modelo de detecção de objetos baseado em transformer de ponta a ponta para nuvens de pontos 3D. Em comparação com os métodos de detecção existentes que empregam vários desvios indutivos específicos para 3D, o 3DETR requer modificações mínimas no bloco transformer vanilla.

Dang *et al*. (2022) propuseram uma estrutura de localização de defeitos de esgoto eficiente e robusta motivado pela arquitetura de última geração do transformador de detecção (DETR), que vê a localização de objetos como um tópico de previsão definido. Uma abordagem de análise da severidade do defeito com base na operação de auto-atenção do transformador para analisar a zona de influência do defeito e o grau do defeito. O sistema proposto superou as abordagens de detecção de objetos padrões anteriores com a maior mAP de 60,2 % no conjunto de dados

coletado.

Egiazarov *et al*. (2020) apresentaram um sistema de detecção de armas baseado em um conjunto de redes neurais convolucionais semânticas que podem ser treinados a baixo custo para detectar apenas componentes específicos de uma arma e que podem ser facilmente agregadas para produzir uma saída robusta. Os resultados mostraram a confiabilidade das redes de componentes individuais e a versatilidade do sistema geral na integração das saídas das redes individuais.

Baldessar (2021) realizou um procedimento automático para detecção de indivíduos portando armas de fogo utilizando rede neural artificial convolucional do tipo YOLOv3 e fazendo uso do dataset de armas de fogo da Universidade de Granada. Com os experimentos realizados a rede foi capaz de realizar detecção de armas de fogo com uma acurácia de 82,3 %, uma precisão de 96,5 % e uma taxa de recall de 84,5 %.

Lim *et al*. (2019) construíram um conjunto de dados que contém armas de uma perspectiva de CFTV como uma solução para a ausência de um conjunto de dados de armas com base em vigilância para aprendizagem de redes neurais profundas. Nesse conjunto de dados, levaram em consideração fatores como condições de iluminação, configurações de CFTV e ambientes que contribuem para a precisão da detecção de armas de uma perspectiva de CFTV para aumentar a diversidade do conjunto de dados. Para validar o conjunto de dados, treinaram usando M2Det como um detector de objeto baseado em Multi-Level Feature Pyramid Network (MLFPN, rede de pirâmide de características multinível, em inglês). Os resultados experimentais indicaram que o conjunto de dados proposto pelos autores aumenta a precisão média de detecção de arma em diferentes escalas em até 18 % em comparação com as abordagens existentes em detecção de armas de fogo.

Verma *et al*. (2017) propuseram uma nova abordagem de detecção de armas de mão para sistemas de alerta e vigilância. O sistema inclui arquitetura VGG-16 baseada em CNN como recurso extrator, seguido por classificadores de última geração implementados em um banco de dados de armas padrão. Os resultados mais promissores obtiveram 93 % de precisão. O sistema pode discernir a existência de inúmeras armas em tempo real e é robusto em toda a variação de afinidade, escala, rotaçãoe fechamento ou oclusão parcial.

No trabalho de Salido *et al*. (2021), foi apresentado um estudo de três modelos de detecção de objetos baseados em CNN (Faster R-CNN, RetinaNet e YOLOv3) - treinado no conjunto de dados MS COCO - aplicado a detecção de revólver em imagens de vigilância por vídeo. Os três objetivos principais do estudo eram para: 1. Comparar o desempenho dos três modelos; 2. Analisar a influência do ajuste fino com uma rede de backbone descongelada /congelada para o modelo RetinaNet; 3. Analisar a melhoria da qualidade de detecção por treinamento de modelo no conjunto de dados com informações de pose - associadas a revólveres - inclusive por um simples método de mesclar as poses do esqueleto nas imagens de entrada. Os resultados destacaram a melhor precisão média (96,36 %) e recall (97,23 %) obtido pelo RetinaNet ajustado com o backbone ResNet-50 descongelado e a melhor precisão (96,23 %) e valores de pontuação F1 (93,36 %) obtidos por YOLOv3 quando foi treinado no conjunto de dados incluindo informações de pose. Esta última arquitetura foi a única que mostrou uma consistente melhoria - cerca de 2 % - quando a informação de pose foi expressamente considerada durante o treinamento.

No trabalho de Olmos *et al*. (2018), foi apresentado um novo sistema de detecção automática de pistola em vídeos apropriados para fins de vigilância e controle. Os autores construíram o conjunto de dados de treinamento guiados pelos resultados de um classificador baseado em VGG-16, para então avaliar o melhor modelo de classificação sob duas abordagens, a abordagem de janela deslizante e a abordagem de proposta de região. Os resultados mais promissores foram obtidos com o modelo baseado no Faster R-CNN, fornecendo zero falsos positivos, 100 % de recall, um alto número de verdadeiros negativos e boa precisão de 84,21 %. O melhor detector mostrou um alto potencial, mesmo em vídeos do youtube de baixa qualidade e fornece resultados muito satisfatórios como sistema de alarme automático. Entre 30 cenas, ele ativa com sucesso o alarme após cinco sucessivos verdadeiros positivos em um intervalo de tempo menor que 0,2 segundos, em 27 cenas.

González *et al*. (2020) apresentaram o comportamento de um detector de objetos Faster R-CNN usando *Feature Pyramid Network* (FPN, rede de pirâmides de características, em inglês) e treinaram usando imagens sintéticas e reais em um CFTV real. Esse trabalho melhora o estado da arte em conjuntos de dados existentes e mostra que o treinamento com dados sintéticos aumenta a precisão, adicionando

mais variedade ao modelo e os dados reais melhoram a detecção de imagens de CFTV. No entanto, os resultados mostraram que há espaço para melhorias em cenários reais, visto que baixa precisão média é obtida e o tempo de inferência também precisa ser aprimorado para se aproximar de um sistema de detecção em tempo real.

Qi *et al*. (2021) optaram por usar ResNet, com diferentes profundidades e compressão no número de canais. Eles treinaram a rede de classificação com 51 mil imagens de armas e 94 mil imagens negativas. O conjunto de teste incluiu 998 amostras positivas e 980 amostras negativas. Os resultados finais do teste mostraram que a ResNet-50 obteve uma acurácia de 97,83 %, um *recall* de 99,69 % e uma precisão de 95,99 %. Até onde se pesquisou, esse trabalho possui a maior detecção de armas e conjunto de dados de classificação disponíveis para fins de pesquisa e desenvolvimento.

Olmos *et al*. (2021) apresentaram um novo sistema, denominado Sistema de Alarme de Nível de Confirmação (MULTICAST) baseado em Redes Neurais Convolucionais (CNN) e Redes de Memória de Longo Prazo do inglês *Long Short Term Memory* (LSTM), que potencializam não só a informação espacial, mas também a informação temporal existente nos vídeos para uma detecção de arma mais confiável. MULTICAST consiste em três etapas, i) o estágio de detecção de arma de mão, ii) um estágio de confirmação espacial baseado em CNN e iii) estágio de confirmação temporal baseado em LSTM. O estágio de confirmação temporal usa as posições da arma detectada em instantes anteriores para prever sua trajetória no próximo quadro. Os experimentos mostraram que MULTICAST reduz em 80 % o número de alarmes falsos em relação ao Faster R-CNN baseado em imagem única detector.

Ruiz-Santaquiteria *et al*. (2021) propuseram um novo método ao combinar, em uma única arquitetura, aparência da arma e informações de pose humana. Primeiro, pontos-chave de pose são estimados para extrair regiões de mão e gerar imagens binárias de pose, que servirão como as entradas do modelo. Então, cada entrada é processada em diferentes sub-redes e combinadas para produzir a caixa delimitadora de arma de fogo. Os resultados obtidos mostram que o combinado modelo melhora o estado da arte da detecção de armas de fogo, alcançando de 4,23 a 18,9 pontos de AP a mais do que a melhor abordagem apresentada em Salido *et al*. (2021).

Ashraf *et al*. (2021) compararam duas arquiteturas para detecção de armas em vídeos de vigilância usando CNN e YOLOv5. Os resultados obtidos mostram uma alta taxa de *recall* e uma alta taxa de velocidade de detecção da arquitetura com YOLOv5. A velocidade atingiu 0,010 s por quadro em comparação com 0,17 s do Faster R-CNN.

Ruprah *et al*. (2023) desenvolveram previsão de crimes baseada na relação pessoa-arma usando técnicas de aprendizado profundo. Nesse trabalho, foram utilizados os modelos como SSD, YOLO e Faster R-CNN para detecção de armas e a biblioteca mediapipe é usada para gerar os pontos de dados do corpo humano e calcular a relação entre as armas com o ser humano. A precisão do R-CNN mais rápido é de 97 %, 93 %, 90 % para armas, facas e pessoas, o que é melhor do que os modelos SSD e YOLO.

Chatterjee *et al.* (2023) desenvolveram uma técnica eficiente de monitoramento de armas de fogo baseada em *deep learning* para a construção de cidades inteligentes seguras. Essa técnica faz uso de um esquema de conjunto (empilhado) usando diferentes técnicas de detecção, como *Faster Region-Based Convolutional Neural Networks* (Faster R-CNN, Redes neurais convolucionais baseadas em regiões mais rápidas, em inglês) e as mais recentes arquiteturas baseadas em EfficientDet para detectar armas e rostos humanos. Esse esquema, melhorou o desempenho de detecção no nível de pós-processamento usando técnicas de supressão não máxima, ponderação não máxima e fusão de caixa ponderada. Os esquemas ensemble obtidos são satisfatórios e melhoram consistentemente em relação aos modelos primários.

Al-Mousa *et al*. (2023) desenvolveram uma arquitetura de sistema de segurança que utiliza técnicas de aprendizagem profunda e de processamento de imagem para a detecção de armas em tempo real. O sistema foi testado e atingiu uma precisão de 92,5 %. Além disso, foi capaz de completar a detecção em apenas 1,6 segundos.

Kambhatla *et al*. (2022) desenvolveram um modelo de detecção de armas de fogo usando *deep learning*. Nesse trabalho, os autores utilizaram o YOLOv5, construindo do zero e coletaram um conjunto de dados de diferentes classes de armas de fogo do Kaggle e do Google Open Images e anotaram manualmente cerca de 2300 amostras. Os resultados atingiram 89,4 %, 70,1 %, 80,5 % de precisão, *recall* e mAP@0.5, respectivamente, e atingiram a maior precisão de 94,1 % por

classe individual.

Considerando os trabalhos analisados, percebe-se que ainda existem espaços para melhorias em sistemas de detecção de armas, seja na questão da velocidade e/ou na questão da precisão. Também, continua sendo um grande desafio projetar um modelo de detecção que supere as variações de forma, aparência e também oclusões da arma de fogo em imagens. Além disso, não foi encontrado na literatura uma proposta que utiliza a abordagem Detection Transformer (DETR) para detecção de armas de fogo em imagens e/ou vídeos. Além do mais, também não foi encontrado o método de múltiplas redes neurais autocoordenadas em sistemas de vigilância. Frente a isso, este trabalho propõe a colaboração de múltiplas redes tendo como base o DETR na detecção de armas de fogo.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado da seguinte forma. Além do atual capítulo, o Capítulo 2 aborda a fundamentação teórica do estudo. Em sequência, o Capítulo 3 apresenta os materiais e os métodos adotados neste trabalho. Logo após, o Capítulo 4 apresenta e discute os resultados obtidos. O trabalho finaliza-se no Capítulo 5 com as conclusões e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 DETECÇÃO DE OBJETOS

Detecção de objeto é uma subárea da visão computacional com o objetivo de reconhecimento de objetos. De acordo com Russakovsky *et al*. (2015), o termo reconhecimento de objetos está relacionado a um conjunto de tarefas como a classificação de imagens, localização de objetos, detecção de objetos, segmentação semântica e segmentação de instância.

A identificação do objeto acontece por determinados recursos característicos (tamanho, proporções, cor, texturas), para curvaturas, bordas, proporções ou modelos de exemplo. A detecção de objetos pode ser aplicada em várias aplicações, como em vídeo de vigilância, robótica, carros autônomos, entre outros.

De modo geral, existem duas categorias em relação aos algoritmos de detecção de objetos: algoritmos em dois estágios e algoritmos em um estágio (Math Works, 2022). Os detectores de um estágio têm altas velocidades de inferência e detectores de dois estágios têm alta precisão de localização e reconhecimento. Nas próximas subseções são abordados, com mais ênfase, tais métodos.

2.2 MÉTODOS DE DETECÇÃO DE OBJETOS

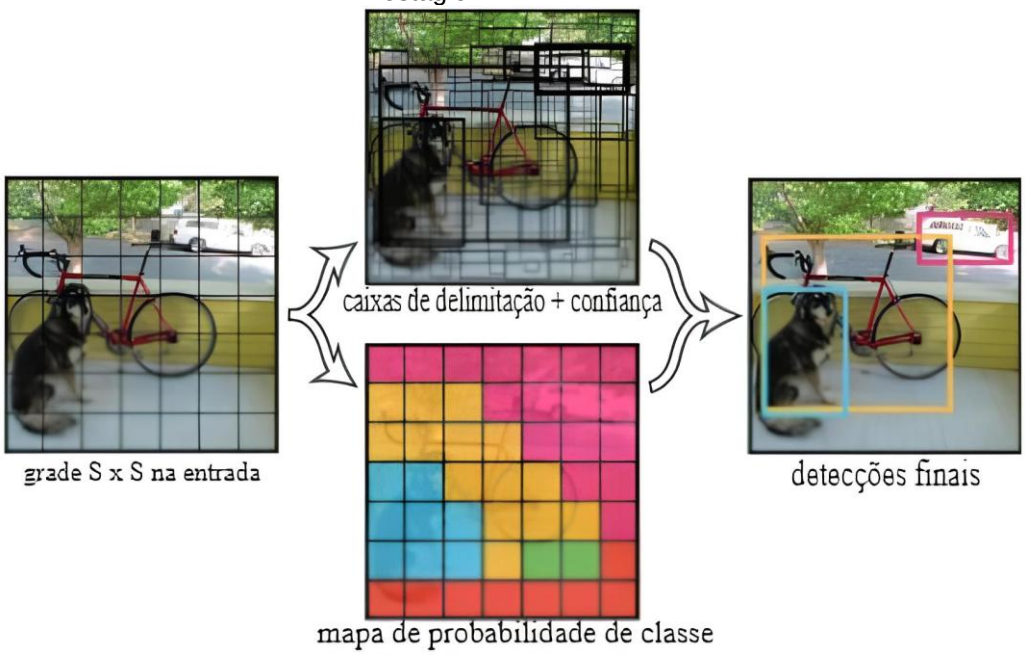
2.2.1 Algoritmos de detecção de objetos em um estágio

Ao contrário dos detectores de dois estágios, não há uma *Region Proposal Network* (RPN, Redes de propostas de regiões, em inglês) nos detectores do tipo um estágio. Esses, por sua vez, tentam realizar tanto a classificação quanto à localização do objeto na imagem de uma só vez a partir do mapa de características. Os exemplos mais comuns de detectores de objetos de um estágio são YOLO, SSD, SqueezeDet e DetectNet.

A seguir, na Figura 1, observa-se como esse tipo de método funciona. A imagem de é tratada como uma grade de quadrados de mesmo tamanho e em cada uma destas partições a rede irá calcular a probabilidade de determinada classe estar ali. Por esta abordagem irá supor que o quadrado que apresenta a maior probabilidade para cada determinada classe é onde se encontra o centro daquele objeto. Sendo assim, a rede irá aplicar a técnica de *non-maximum supression* em

todas as caixas de seleção no arredor desta partição retornando a área em que acredita estar o objeto.

Figura 1 - Procedimento realizado por um detector de objetos de um estágio.



Fonte: adaptada de Readmon *et al*. (2016).

Em função de sua simplicidade, os detectores de um estágio acabam por ser mais velozes do que os detectores do outro tipo, realizando a tarefa de detecção de objetos em um tempo expressivamente menor. Contudo, eles acabam sofrendo no quesito exatidão, uma vez que costumam mostrar-se inferiores aos métodos com RPN.

As redes de um estágio, por sua arquitetura, acabam tendo que lidar com uma quantidade muito maior de proposições de localização de objetos e, embora existam métodos para tentar contornar o problema, eles são ineficientes. Isso ocorre uma vez que, para o treinamento, exemplos facilmente classificáveis de plano de fundo acabam por exercer domínio.

A solução encontrada pelos autores foi o desenvolvimento de uma nova função de custo inspirada na *cross entropy*, a qual recebeu o nome de *focal loss*. Essa função foca o treinamento em um conjunto esparso de exemplos difíceis de classificar, o qual previne a ocorrência de um vasto número de negativos facilmente classificáveis que acabam por sobrecarregar o detector durante o treinamento (LIN; GOYAL *et al*., 2017).

2.2.2 Algoritmos de detecção de objetos em dois estágios

Os detectores de dois estágios assumem a liderança em precisão de detecção. Nesses detectores, as regiões aproximadas do objeto são propostas usando recursos profundos antes que esses recursos sejam usados para a classificação, bem como a regressão de caixa delimitadora para o objeto candidato.

Segundo Boesch (2024), vários detectores de dois estágios incluem rede neural convolucional de região (R-CNN), com as evoluções Faster R-CNN e Mask R-CNN. A última evolução é chamada de R-CNN granulado (G-RCNN). Esses detectores atingem a maior precisão de detecção, mas geralmente são mais lentos. Devido às muitas etapas de inferência por imagem, o desempenho (quadros por segundo) não é tão bom quanto os detectores de um estágio. Além do mais, geralmente não são treináveis de ponta a ponta porque o corte é uma operação não diferenciável.

2.3 TRANSFORMERS

Rede transformadora, ou *transformer*, é uma arquitetura de rede neural artificial proposta pela primeira vez no artigo *Attention is all you need* e foi rapidamente estabelecido como a arquitetura líder para a maioria dos aplicativos de dados de texto.

O *Transformer* é uma arquitetura que usa atenção para melhorar significativamente o desempenho dos modelos de tradução PLN de aprendizado profundo. O GPT-4, indiscutivelmente o modelo de PLN mais avançado até hoje, é capaz de realizar uma ampla variedade de tarefas de PLN de maneira semelhante à humana.

Os transformadores são uma arquitetura de modelo de aprendizado de máquina, como redes neurais de memória de curto prazo (LSTMs) e redes neurais convolucionais (CNNs). Uma diferença fundamental entre os transformadores e outras arquiteturas é que os transformadores são altamente paralelizáveis, tornando-os muito otimizados para computação, o que permite treinar modelos extremamente grandes. O segundo ponto que diferencia os transformadores é o uso inteligente da atenção.

O mecanismo de atenção foi usado pela primeira vez em 2014 na visão computacional para tentar entender o que uma rede neural está olhando enquanto faz uma previsão. Esse foi um dos primeiros passos para tentar entender as saídas das CNNs. Em 2015, a atenção foi usado primeiro em PLN em tradução automática alinhada. Por fim, em 2017, o mecanismo de atenção foi utilizado nas redes *Transformer* para modelagem de linguagem. Desde então, os transformadores ultrapassaram as precisões de previsão das Redes Neurais Recorrentes (RNNs), para tornarem-se o estado da arte para tarefas de PLN.

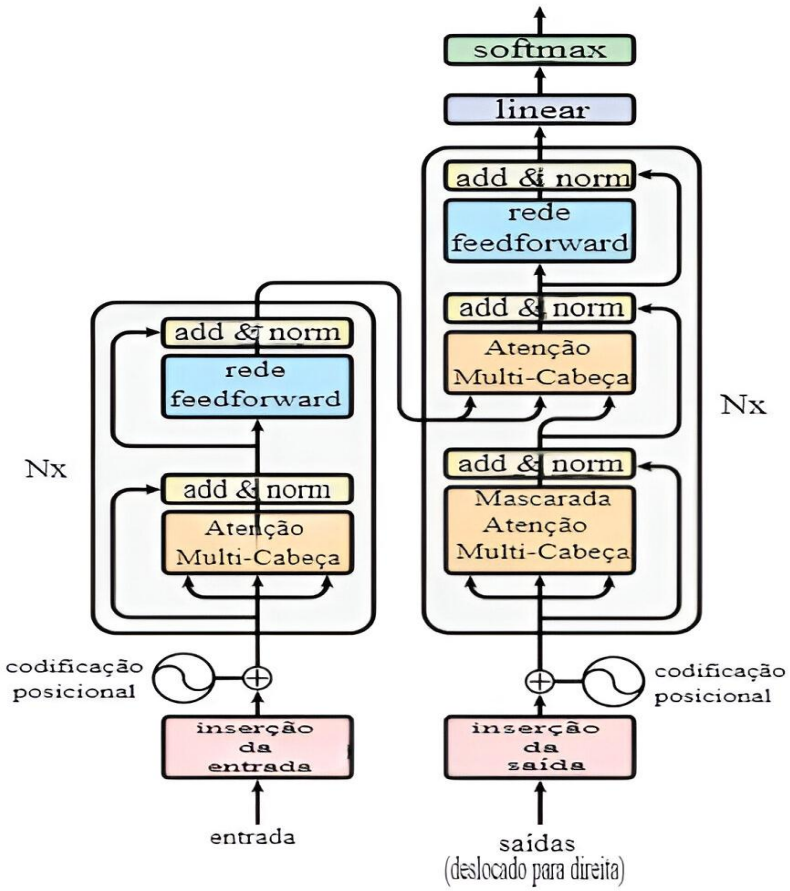
A arquitetura do *Transformer* destaca-se no tratamento de dados de texto que são inerentemente sequenciais. Em seu núcleo, ele contém uma pilha de camadas do codificador e camadas do decodificador. A pilha do codificador e a pilha do decodificador têm, cada uma, suas camadas de incorporação correspondentes para suas respectivas entradas. Finalmente, há uma camada de saída para gerar a saída final. Na Figura 2, é possível observar sua arquitetura.

Transformers são muito versáteis e são usados para a maioria das tarefas de PLN, como modelos de linguagem e classificação de texto. Eles são frequentemente usados em modelos de sequência a sequência para aplicativos como tradução automática, resumo de texto, respostas e perguntas, reconhecimento de entidade nomeada e reconhecimento de fala.

Existem diferentes tipos de arquitetura do *transformer* para diferentes problemas. A camada codificadora básica é usada como um bloco de construção comum para essas arquiteturas, com diferentes 'cabeças' específicas do aplicativo, dependendo do problema que está sendo resolvido.

Em 2020, uma nova técnica foi introduzida pela equipe de pesquisa do Facebook em um artigo chamado *End-to-End Object Detection with Transformers (CARION et al. ,2020)*. Nesse trabalho, foi apresentado um novo modelo de detecção de objetos fazendo uso da arquitetura *transformers*, que até então era empregado em soluções de PLN. No próximo subcapítulo, essa arquitetura é abordada com mais detalhes sobre.

Figura 2 - Arquitetura do transformador



Fonte: adaptada de Vaswani, *et al*. (2017).

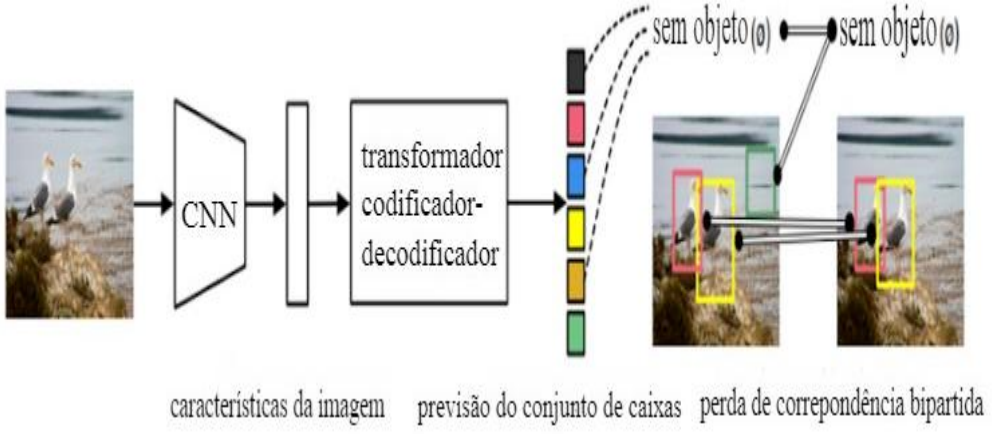
2.3.1 Detection transformers (DETR)

O Transformador de Detecção ou DETR foi apresentado por Carion *et al*. (2020). Eles usaram uma combinação de CNN e transformador para criar um detector treinável de ponta a ponta e removeu qualquer módulo artesanal, como a supressão não máxima ou geração de âncoras. A imagem de entrada é primeiro passada através de uma rede de backbone CNN para obter características da imagem e depois através de um módulo transformador codificador-decodificador que emite diretamente as previsões da caixa delimitadora do objeto. Para combinar essas previsões com os rótulos de verdade, uma perda de correspondência bipartida é usada. Essa correspondência bipartida é feita usando o chamado algoritmo húngaro.

A ideia de usar um transformador é que ele pode usar seu mecanismo de auto-atenção para livrar- se de previsões duplicadas. A abordagem do transformador aqui usa um transformador paralelo, ou seja, todos os objetos são previstos em paralelo,

em vez de ter uma previsão sequencial, como geralmente é feito no processamento de linguagem natural, onde a geração da próxima palavra influencia a palavra gerada posteriormente.

Figura 3 - Arquitetura DETR de alto nível

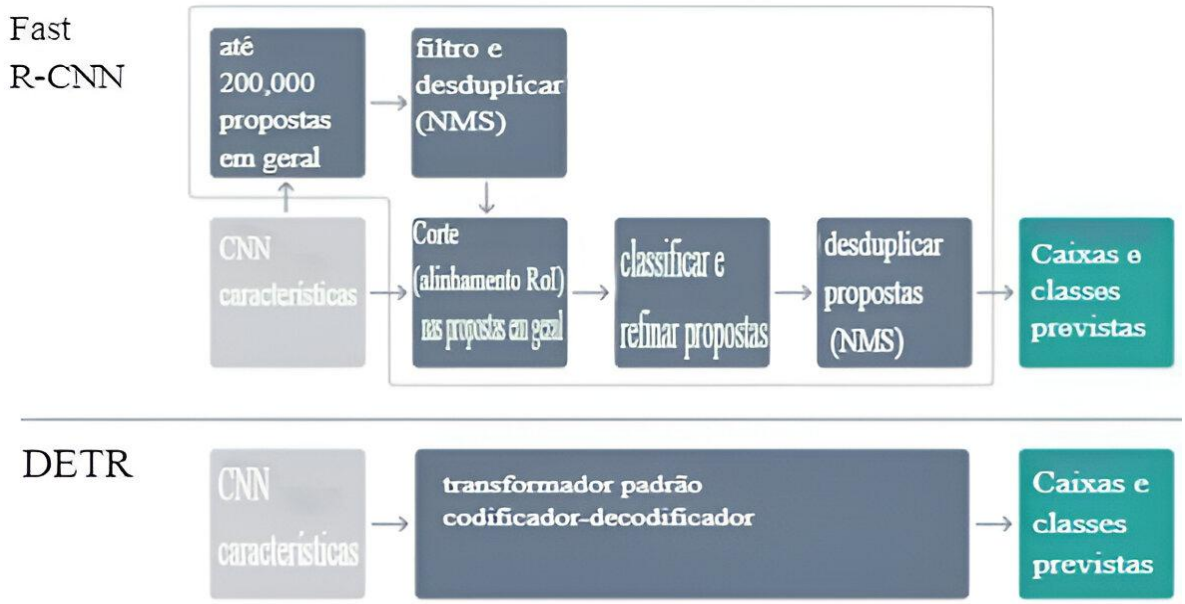


Fonte: adaptada de Carion *et al*. (2020).

O DETR combina o desempenho de métodos de última geração, como a linha de base Faster R- CNN bem estabelecida e altamente otimizada no desafiador conjunto de dados de detecção de objetos COCO, ao mesmo tempo em que simplifica e agiliza bastante a arquitetura. Na Figura 4, é possível visualizar a simplicidade do DETR em relação ao Faster R-CNN.

Os sistemas tradicionais de detecção de dois estágios, como Faster R-CNN, preveem caixas delimitadoras de objetos filtrando um grande número de regiões candidatas, que geralmente são uma função dos recursos da CNN. Cada região selecionada é usada em uma etapa de refinamento, que envolve o recorte dos recursos da CNN no local definido pela região, classificando cada região de forma independente e refinando sua localização. Finalmente, uma etapa de supressão não máxima é aplicada para remover caixas duplicadas. O DETR simplifica o *pipeline* de detecção aproveitando uma arquitetura padrão do transformador para executar as operações quesão tradicionalmente específicas para detecção de objetos.

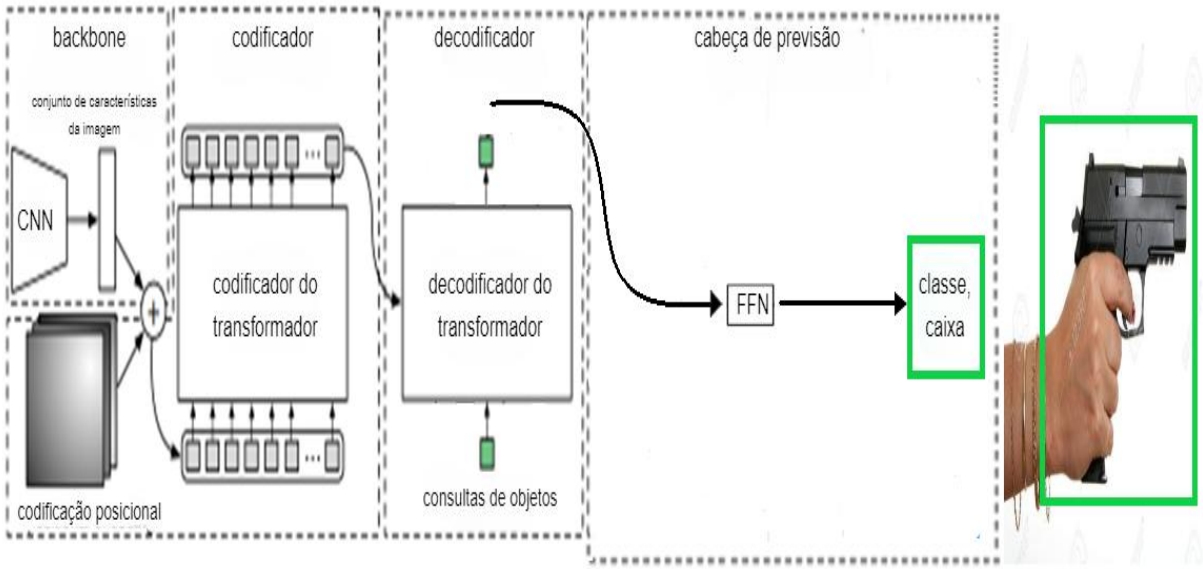
Figura 4 - Comparação das arquiteturas Faster R-CNN e DETR



Fonte: adaptada de Akdogan (2022).

A arquitetura geral do DETR contém três componentes principais: um backbone CNN para extrair uma representação compacta de características, um transformador codificador-decodificador e uma rede *Feed Forward* (FFN) que faz a previsão final de detecção. Na Figura 5, é possível observar aarquitetura do DETR.

Figura 5 - Detalhes da arquitetura DETR.



Fonte: adaptada de Carion *et al*. (2020).

Backbone

Na arquitetura DETR (Detection Transformer), os backbones desempenham um papel crucial na extração de características das imagens. Eles são responsáveis por gerar uma representação de alta qualidade da imagem, que é então usada pelo transformador para realizar a detecção de objetos (CARION *et al.*, 2020).

Os *backbones* são redes neurais convolucionais (CNNs) que funcionam como extratores de características. Eles convertem a imagem original em um conjunto de mapas de características de menor dimensão, mas com informações detalhadas sobre a presença e a localização dos objetos (SHAH, 2020).

Entre as arquiteturas de backbone mais frequentemente utilizadas com o DETR, encontra-se a ResNet (*Residual Networks*). De acordo com Hasanah *et al*. (2023), a ResNet, apresenta diferentes variações em suas arquiteturas, embora todas se originem da mesma base. A principal diferença entre essas arquiteturas é o número de camadas e parâmetros, que são ajustados conforme o modelo. Por exemplo, a ResNet-50 possui 50 camadas, a ResNet-101 tem 101 camadas e a ResNet-152 conta com 152 camadas.

A ResNet-18 é uma versão mais leve e rápida, sendo composta por 17 camadas convolucionais, uma camada de *pooling* máximo com filtro de 3 x 3 e uma camada totalmente conectada. Este modelo clássico, possui 33,16 milhões de parâmetros, e a função de ativação ReLU juntamente com a *batch normalization* (BN, normalização de lote, em inglês) são aplicadas após cada bloco básico de camadas convolucionais (GAO *et al*., 2021). A estrutura do ResNet-18 pode ser vista na Tabela 1.

A escolha do backbone afeta a precisão e a velocidade do modelo de detecção, com backbones mais profundos geralmente fornecendo representações mais ricas, mas exigindo mais recursos computacionais. Além da ResNet, outras redes como EfficientNet, ViT (*Vision Transformer*) e DenseNet também podem ser utilizadas como backbones no DETR, cada uma trazendo suas próprias vantagens em termos de eficiência, precisão e complexidade.

Tabela 1 – Estrutura da ResNet-18

Nome da camada	Tamanho da saída	18 camadas
Conv1	112 × 112	7 × 7, 64, stride 2
Conv2_x	56 × 56	3 × 3 max pool, stride 2 $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
Conv3_x	28 × 28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
Conv4_x	14 × 14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
Conv5_x	7 × 7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$

Fonte: Gao *et al*. (2021).

Codificador do transformador

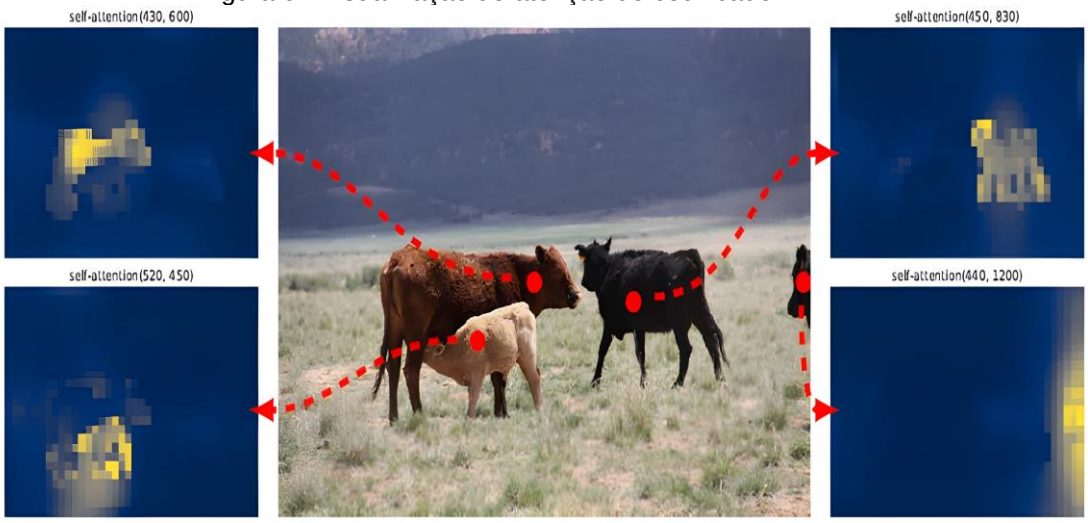
O codificador do transformador codifica estas caraterísticas numa sequência de vetores de caraterísticas. Como o codificador contém blocos de auto-atenção com várias cabeças, permite captar informações contextuais através de dependências de longo alcance entre diferentes partes de uma imagem (SHASTRI, 2024). Nesta fase, outro elemento crucial do *pipeline* é a adição de codificações posicionais à saída da CNN. Como os transformadores não têm inerentemente uma compreensão espacial, estas codificações posicionais informam sobre as posições relativas dos objetos na imagem.

O codificador é composto por uma pilha de seis camadas idênticas (Nx = 6 na Fig. 2). Cada camada tem duas subcamadas. A primeira é um mecanismo de auto-atenção multicabeça e a segunda é uma rede *feed-forward* simples, totalmente conectada em termos de posição (CARION *et al*., 2020).

Dessa forma, o codificador desenvolve nessa etapa a capacidade de distinguir objetos e fundos, podendo até distinguir diferentes objetos da mesma classe. Segundo Carion *et al*. (2020), graças a isso, o codificador realiza apenas tarefas simples. Na figura a seguir, é apresentada a visualização de atenção do

codificador.

Figura 6 - Visualização de atenção do codificador.



Fonte: Carion *et al*. (2020).

Decodificador do transformador

O descodificador do transformador aprende as relações entre as caraterísticas codificadas pela CNN e as consultas de objetos aprendíveis. Normalmente, são necessárias consultas, chaves e valores para calcular a auto-atenção na PNL. Por isso, na visão computacional, o DETR introduz o conceito de consultas de objetos, que se referem às representações aprendidas dos objetos que o modelo precisa prever (SHAH, 2020). O número de consultas de objetos é pré-determinado e permanece fixo. As chaves indicam as localizações espaciais na imagem, enquanto os valores contêm informações sobre as caraterísticas.

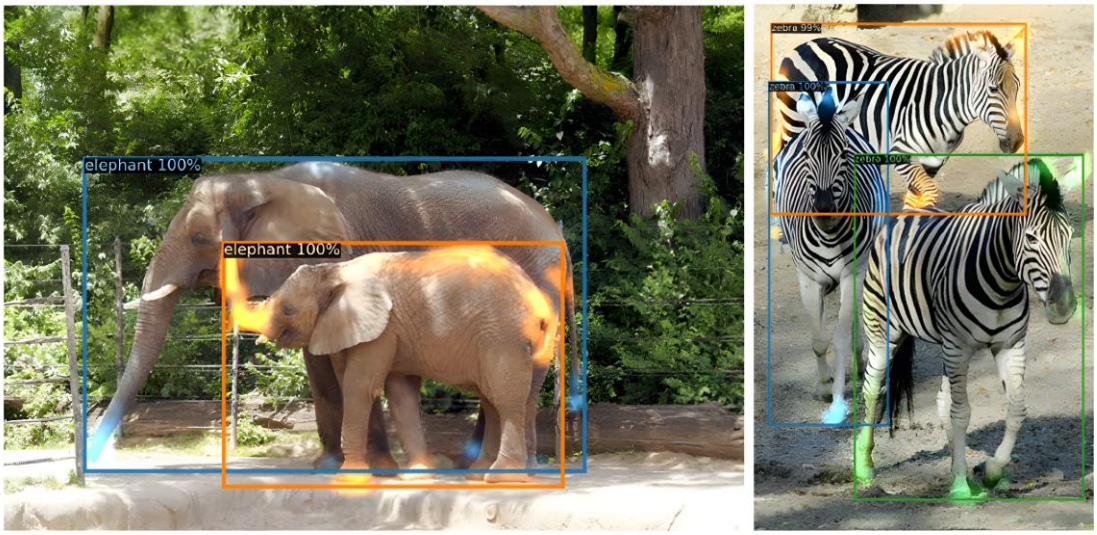
O decodificador também é composto por uma pilha de seis camadas idênticas. Além das duas subcamadas em cada camada do codificador, o decodificador insere uma terceira subcamada, que executa a atenção multicabeça sobre a saída da pilha do codificador (CARION *et al*., 2020).

Semelhante ao codificador, conexões residuais são empregadas em torno de cada uma das subcamadas, seguidas pela normalização da camada. A subcamada de auto-atenção na pilha do decodificador é modificada para evitar que posições atendam a posições, futuras, subsequentes.

Assim sendo, o decodificador costuma concentrar-se nas extremidades do objeto. Ao observar a Figura 7, percebe-se que, em vez de apenas se concentrar em

todas as bordas do pixel do objeto, uma região que realmente precisa ser ativada para mostrar a área da caixa delimitadora é a que recebe mais ativação. Isso significa que, ao invés de tratar todas as bordas de maneira uniforme, o decodificador prioriza as partes do objeto que são mais relevantes para a definição da caixa delimitadora.

Figura 7 - Visualização de atenção do decodificador.



Fonte: Carion *et al*. (2020).

Cabeça de previsão

As cabeças de previsão geram as caixas delimitadoras e as categorias dos objetos identificados. Elas são formadas por cabeças de rede feed-forward que estimam tanto a caixa delimitadora quanto a classe dos objetos detectados, ou uma classe de “sem objeto” quando não há detecções (CARION *et al*., 2020). Além disso, o DETR aplica a técnica de correspondência bipartida para assegurar que as caixas delimitadoras previstas correspondem aos objetos reais. Esse método também contribui para aprimorar o treinamento do modelo (SHASTRI, 2024).

Essa abordagem de correspondência bipartida é crucial, pois permite que o modelo aprenda de maneira mais eficaz, alinhando a precisão com as verdadeiras instâncias dos objetos. Ao fazer isso, o DETR não apenas melhora a precisão das detecções, mas também facilita a generalização do modelo em diferentes cenários. Com a combinação das cabeças de previsão e a técnica de correspondência, o sistema se torna mais robusto e capaz de lidar com uma variedade de desafios em tarefas de detecção de objetos (HUDA, 2023).

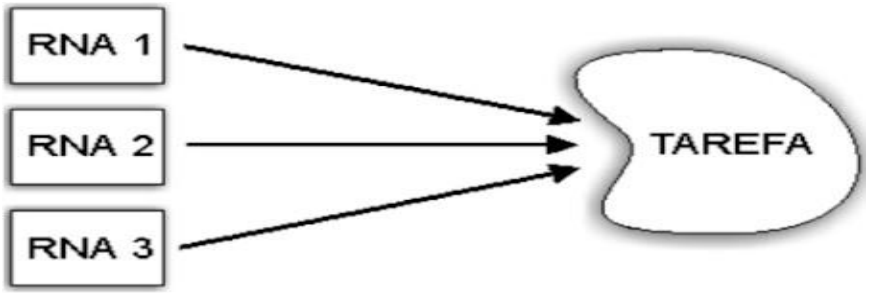
Em geral, o DETR oferece uma nova forma de detecção de objetos de ponta a ponta. A consulta do objeto aprende gradualmente características de instância durante a interação com as características da imagem. A correspondência bipartida permite uma previsão direta para adaptar-se facilmente à tarefa de atribuição de rótulos um a um e assim eliminar o pós-processamento tradicional.

2.4 MÚLTIPLAS REDES NEURAIS

Múltiplas Redes Neurais (MRN) consistem em um arranjo de várias redes neurais, normalmente do mesmo tipo, solucionando um mesmo problema. A vantagem de usar múltiplas RNAs é a possibilidade de uma rede aprender parte da solução e assim as demais redes fazem o mesmo, desta forma o conjunto aprende a resolver o problema todo. Segundo Yang *et al*. (2001), a aplicação desta técnica apresenta melhoria na robustez, capacidade de predição e generalização do modelo neural.

Ao usar múltiplas RNAs, existem formas de empregar as várias redes: combinação em conjunto, combinação modular e uma junção das duas formas anteriores. A combinação em conjunto, também chamada de ensemble, representa uma junção de redes, onde cada uma tenta resolver o problema inteiro. Perrone *et al*. (1993) sugerem que as RNAs componentes do ensemble devam ter diferentes arquiteturas, devam ser treinadas com algoritmos diferentes e também com conjuntos de dados de diferentes. A Figura 8 representa esse tipo de combinação.

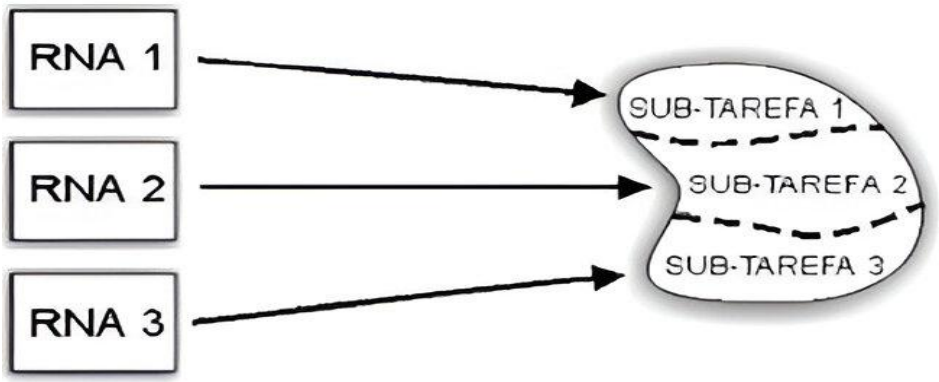
Figura 8 - Combinação do tipo *ensemble*



Fonte: de Almeida Neto (2003).

Já na combinação do tipo modular, o problema é dividido em partes e cada rede tentará resolver cada parte isoladamente. A Figura 9 faz a representação de uma combinação do tipo modular.

Figura 9 - Combinação do tipo modular



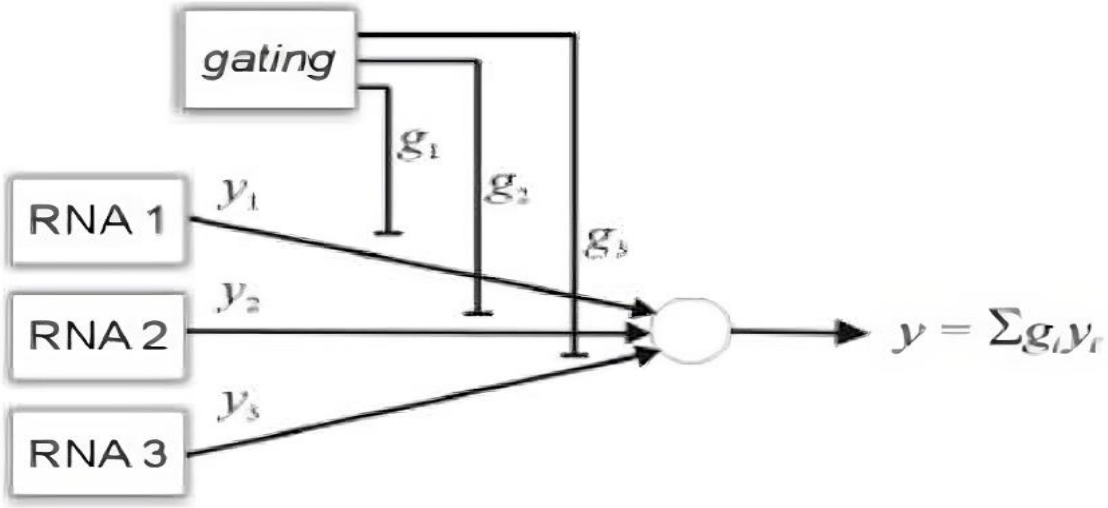
Fonte: de Almeida Neto (2003).

Há também a possibilidade de unir características dos dois tipos de combinação, onde o problema é dividido em subproblemas e várias redes tentam resolver cada um desses subproblemas. Neste trabalho, foi utilizada a combinação do tipo *ensemble*.

Porém, ao se utilizar mais de uma rede na resolução de um problema, torna-se necessária a utilização de um mecanismo que faça com que as redes colaborem entre si, evitando que o aprendizado de uma rede seja prejudicado à medida que outra rede efetua seu treinamento. Para isto, muitos projetistas utilizam um coordenador para evitar conflitos no aprendizado das redes.

Quanto à sua implementação, um coordenador pode apresentar-se como um modelo lógico- matemático, como pode ser visto nos trabalhos de Hashem (1994; 1995 e 1997), onde a saída das RNA é processada para formar uma única saída através de uma combinação linear ótima. Mas também podem-se empregar técnicas de IA para realizar a implementação do coordenador, entre elas, tem-se o uso de sistemas fuzzy, como pode ser visto em Cho *et al*. (1995), algoritmo genético, utilizado em Cho (2002), ou mesmo outra RNA, chamada de *gating*, como pode ser visualizado na Figura 10. Torna-se, portanto, uma responsabilidade a mais ao projetista ter que projetar e implementar mais um elemento, além das RNAs, que pode muitas vezes torna-se uma tarefa muito custosa, podendo comprometer o trabalho como um todo.

Figura 10 - MRNA autocoordenadas com o coordenador do tipo gating



Fonte: de Almeida Neto (2003).

2.4.1 Múltiplas redes neurais autocoordenadas

O conceito de MRNA autocoordenadas consiste na utilização de mais de uma rede para resolver um determinado problema (DE ALMEIDA NETO *et al*., 2010). Métodos parecidos estão presentes na literatura, porém o método desenvolvido por de Almeida Neto *et al*. (2010) propõe um diferencial, na qual a inserção de redes ocorre de maneira sequencial, evitando conflitos entre as redes e dispensando o uso de um coordenador.

A colaboração entre as redes sem um coordenador acontece na definição do valor desejado a ser aprendido por cada rede. Assim, ao ser treinada, uma rede tem como objetivo apresentar uma saída que, ao ser somada com a saída das outras redes, iguale ao valor desejado para a solução do problema.

Em termos matemáticos, o valor de erro utilizado no cálculo da atualização dos pesos da n-ésima rede é determinado pelo valor desejado para a solução do problema subtraído das saídas de todas as redes já anteriormente treinadas. A Equação 1 a seguir apresenta a formulação desse cálculo:

$$e = d - y_1 - y_2 \ldots - y_n \quad (1)$$

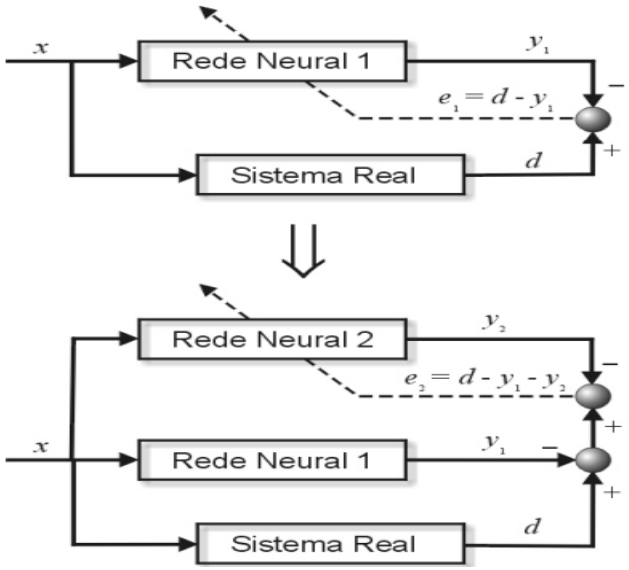
sendo e o erro da n-ésima rede, d o valor desejado para a solução do problema, y_1 a saída da rede 1 e assim sucessivamente até a rede *n* com saída igual ao y_1 anterior. Na Figura 11, é apresentado esse esquema de inserção de novas redes.

Todavia para que a colaboração entre as RNAs funcione corretamente, é preciso garantir que toda nova rede adicionada ao conjunto deve começar com uma saída igual a zero (DE ALMEIDA NETO, 2003). Isso é importante porque a nova rede deve iniciar seu treinamento a partir do ponto em que as redes existentes pararam, assegurando que a saída total do sistema não mude após a adição da nova rede.

Assim, ao assegurar que a saída da nova rede seja zero ao ser adicionada, evita-se alterações indesejadas no resultado global. Essa abordagem permite que o impacto da nova rede seja neutro no início, facilitando a integração sem alterar o comportamento do sistema como um todo. Dessa forma, o somatório das saídas das redes após a inserção permanece igual ao que era antes da adição da nova rede (DE ALMEIDA NETO, 2003).

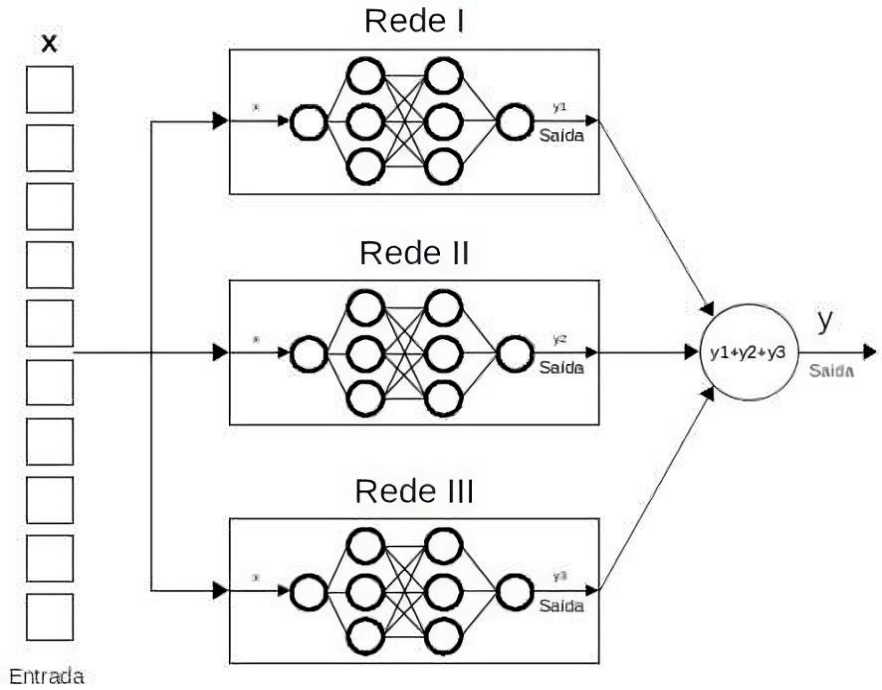
Isto posto, o método citado consiste em um modelo de treinamento em que as redes são inseridas em sequência. Começando com apenas uma rede, esta é treinada e assim que esgota sua capacidade de aprender (erro estabilizado), seus pesos não são mais ajustados (parada no treinamento da primeira rede), então uma segunda rede é inserida no grupo e colocada em treinamento com a finalidade de adquirir o conhecimento não obtido pela primeira. A saída do modelo proposto é o somatório das saídas de todas as redes. A Figura 12 ilustra o esquema do modelo sem coordenador.

Figura 11 – Inserção de novas redes sem o coordenador



Fonte: de Almeida Neto (2003).

Figura 12 - Ilustração do funcionamento MRNA autocoordenadas



Fonte: Teixeira (2017).

Inspirando-se nessa abordagem, a ideia deste projeto é utilizar a técnica de múltiplas redes neurais autocoordenadas na parte *full connected* da rede DETR, para reconhecimento de objetos em tempo real. Essa utilização visa a melhorar a precisão da rede DETR no reconhecimento de objetos pequenos, permitindo que esta se adapte dinamicamente a diferentes condições de iluminação, ângulos e contextos.

Além disso, a utilização de redes autocoordenadas possibilita uma maior eficiência no projeto. Ao se usar apenas uma rede rasa na parte *full connected*, há muitas configurações dessa rede rasa a serem testadas, até que se encontre uma que consiga aprender a classificação correta. Isso leva muito tempo. Já com o uso de múltiplas RNAs autocoordenadas, a quantidade de configurações a testar é drasticamente reduzida. Pode-se relaxar na busca de uma configuração de rede ideal usando uma configuração de hiperparâmetros qualquer, pois a inserção de uma segunda rede vai melhorar o desempenho aprendendo o que faltou ser aprendido pela rede anterior. E a segunda rede também não precisa ter uma configuração ideal, pois uma terceira rede pode ser inserida para aprender o que faltou ser aprendido pelas redes anteriores. Assim, pode-se inserir sempre uma rede sem precisar ter uma configuração otimizada.

3 METODOLOGIA

3.1 MATERIAIS

Considerando o problema de detecção de armas em câmeras de vigilância por vídeo, foi utilizado um conjunto de dados o mais próximo possível de imagens encontradas em imagens de vídeos de segurança. A base de dados de imagens chamada Monash Guns Dataset de Lim *et al*. (2021) tem armas de uma forma representativa e também com uma variedade de armas diferentes. Os autores descrevem o conjunto de dados como uma mistura de armas canônicas e imagens não canônicas de armas gravadas semelhantes a câmeras de vigilância por vídeo.

A diversidade das imagens no Monash Guns Dataset ajuda a melhorar a precisão dos modelos ao lidar com diferentes condições de iluminação, ângulos e contextos em que as armas podem aparecer. O *dataset* inclui imagens com diferentes níveis de qualidade, abrangendo as de alta resolução, que permitem uma análise detalhada dos detalhes e características das armas. Na Figura 13, é possível observar um exemplo de uma imagem do conjunto de dados com boa qualidade e luminosidade.

Figura 13 - Exemplo de uma imagem CFTV de boa qualidade

Fonte: Lim *et al*. (2021).

Essa base também possui imagens de baixa qualidade, que são úteis para testar a robustez de modelos de reconhecimento em condições menos ideais. Essa diversidade na qualidade das imagens permite que os pesquisadores desenvolvam e avaliem algoritmos em uma ampla gama de cenários realistas. Na Figura 14, mostra-se um exemplo de uma imagem de baixa qualidade.

Figura 14 - Exemplo de uma imagem CFTV de baixa qualidade

Fonte: Lim *et al*. (2021).

O Monash Guns Dataset contém um total de 8000 imagens com dimensões de 512 x 512 pixels. Entre essas imagens, há 2.857 imagens extraídas de 250 vídeos de CFTV e o restante adquiridas da internet. O acesso ao *dataset* é facilitado por meio do repositório no GitHub disponível no endereço: http://github.com/MarcusLimJunYi/Monash-Guns-Dataset/tree/master?tab=readme-ov-file. Este conjunto de dados é estruturado para suportar o treinamento e a avaliação de modelos de reconhecimento de armas, oferecendo uma quantidade considerável de imagens com uma resolução adequada para tarefas de detecção e classificação.

Além da base de dados, é importante mencionar que para o desenvolvimento deste trabalho, foi utilizada a linguagem de programação Python, empregando a biblioteca PyTorch, que é amplamente reconhecida por suas capacidades em aprendizado profundo e redes neurais. PyTorch foi escolhido devido à sua flexibilidade e eficiência na construção e treinamento de modelos complexos. Além disso, foram integradas outras ferramentas e bibliotecas, como NumPy e pandas, para manipulação e análise de dados, e matplotlib para visualização de resultados.

O ambiente de desenvolvimento utilizado para este trabalho é configurado com especificações robustas para garantir um desempenho eficiente e confiável. O sistema é equipado com 31,3 GB de RAM, proporcionando amplo espaço para

operações multitarefa e processamento de dados intensivos. O processador Intel® Core™ i7-10700F, com 16 núcleos e uma velocidade base de 2,90 GHz, oferece um desempenho de processamento potente, ideal para tarefas computacionais complexas. A placa gráfica NV137 complementa o sistema com suporte a gráficos avançados, enquanto a capacidade de disco de 1,5 TB assegura espaço suficiente para armazenamento de grandes volumes de dados e projetos. O sistema operacional usado foi Ubuntu 20.04.4 LTS, na versão de 64 bits, garantindo uma base estável e atualizada para o desenvolvimento. Essa combinação de ferramentas técnicas permitiu uma abordagem robusta e eficaz para o processamento e análise das imagens no contexto do projeto.

3.2 MÉTODOS

Antes de realizar o treinamento, a base passou por alguns pré-processamentos, como a remoção de dados duplicados ou dados incompletos. O ajuste realizado permitiu a adaptação ao modelo e a redução no tempo de treinamento. Além dos ajustes, a base também passou por uma divisão para uso no treinamento. Essas amostras da base foram divididas entre treino, validação e teste, seguindo respectivamente, as proporções: 80 %, 10 % e 10 %. Desta forma, foram utilizadas 6400 imagens para o treinamento, 800 imagens para validação e esta mesma quantidade para o teste do treinamento.

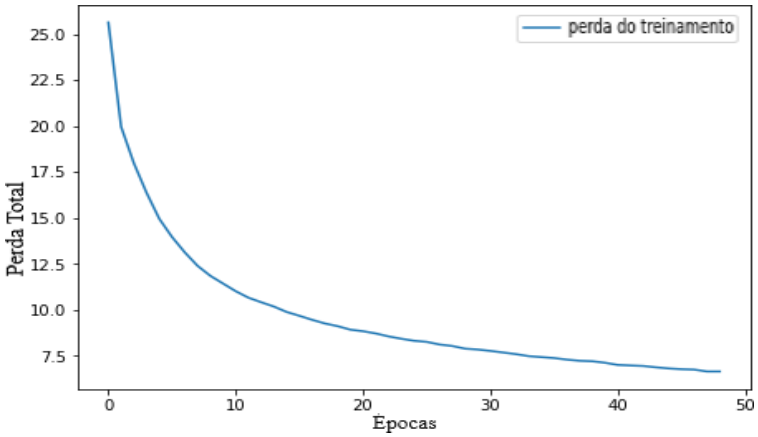
Finalizado a preparação da base, iniciou-se o desenvolvimento da rede. A seguir, será abordado em relação à análise e a implementação da rede utilizada, dividindo em duas fases. Na primeira fase, explora-se a metodologia da rede principal, detalhando os processos que fundamentam seu funcionamento. Em seguida, na segunda fase, foca-se nas redes adicionais, que complementam e ampliam as capacidades da rede principal, discutindo suas especificidades, integrações e contribuições para o desempenho geral do sistema.

3.2.1 Primeira fase - rede principal

A primeira fase do treinamento concentrou-se na busca de um conjunto de hiperparâmetros de uma rede DETR com robustez em reconhecimento e classificação de armas de fogo. Levando-se em consideração que o tempo de treinamento com redes DETR é custoso, não se treinou a rede por muitas épocas.

Na Figura 15, apresenta-se um gráfico do erro durante o treinamento nessa primeira fase em que a rede foi treinada por 47 épocas.

Figura 15 - Primeira fase do treinamento da DETR



Fonte: Autor.

Por fim, a dinâmica de escolha dos hiperparâmetros da primeira fase do treinamento consistiu no tipo de compressão das imagens, no número de camadas da MLP, taxa de aprendizado, número de neurônios e funções de ativação de cada camada. Após a seleção de um conjunto de hiperparâmetros como: 10^{e-5} na taxa de aprendizado, ResNet-18 como *backbone*, seis camadas no codificador e seis camadas no decodificador do transformador, o uso da função sigmoid na última camada, entre outros, deu-se início à fase de treinamento propriamente dita. Essa fase teve o intuito de fazer a rede aprender a detectar. Na primeira fase, o intento era achar um conjunto adequado de hiperparâmetros.

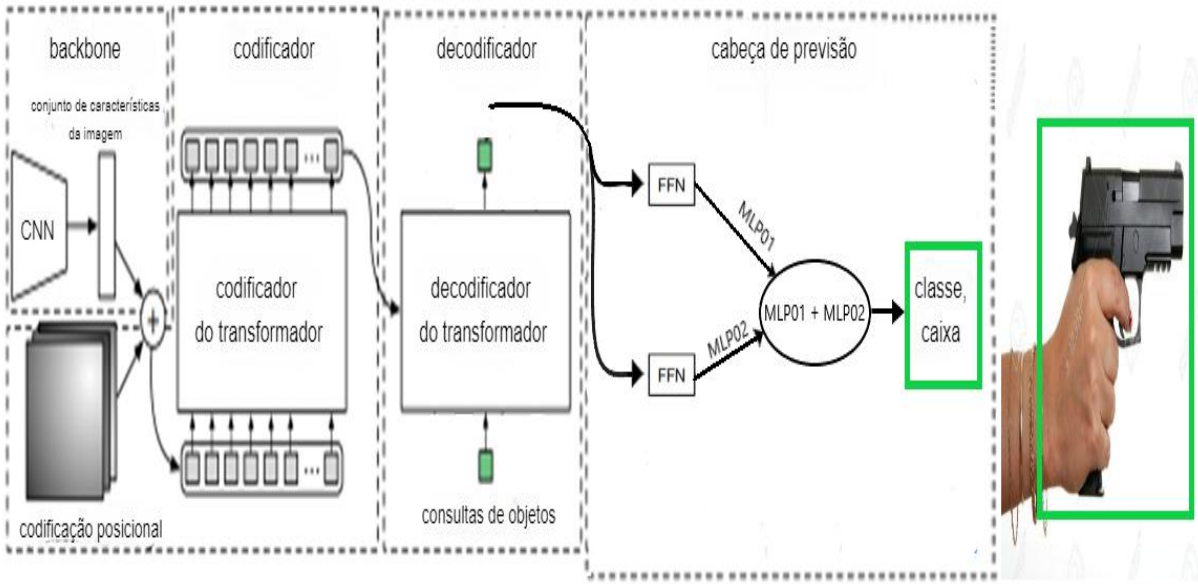
Cada rede DETR configurada foi avaliada para verificação se o aprendizado foi a contento. A avaliação ocorreu com imagens separadas para teste. Nessa avaliação, utilizaram-se as métricas: precisão (Pr), sensibilidade (Se) e F1 Score (F1). Da mesma maneira, com o objetivo de mostrar que as redes adicionais são capazes de serem aplicadas às redes em produção, buscou-se encontrar uma rede com valores próximos do estado da arte e então aperfeiçoá-la com as redes adicionais. Portanto, enquanto não se obtiveram resultados aproximados aos encontrados na literatura, os hiperparâmetros foram modificados e o treinamento reiniciado.

3.2.2 Segunda fase - redes adicionais

Uma vez que acabadas as tentativas de melhorar a rede principal, inicia-se a inserção das redes adicionais. Apesar da Figura 13 apresentar o modelo com uma rede adicional, esta abordagem não está limitada a esse número.

A contribuição deste trabalho começa na proposta de inserção das redes adicionais. Após o treinamento da rede DETR usando apenas uma rede MLP na parte *full connected*, caso uma solução ótima não seja encontrada, o modelo deve ser retreinado. Nesta abordagem, o próximo passo é tentar melhorar a detecção da rede partindo para adição de MLPs auxiliares. A Figura 16 apresenta como fica o modelo com esta abordagem.

Figura 16 - Rede DETR com uma rede adicional



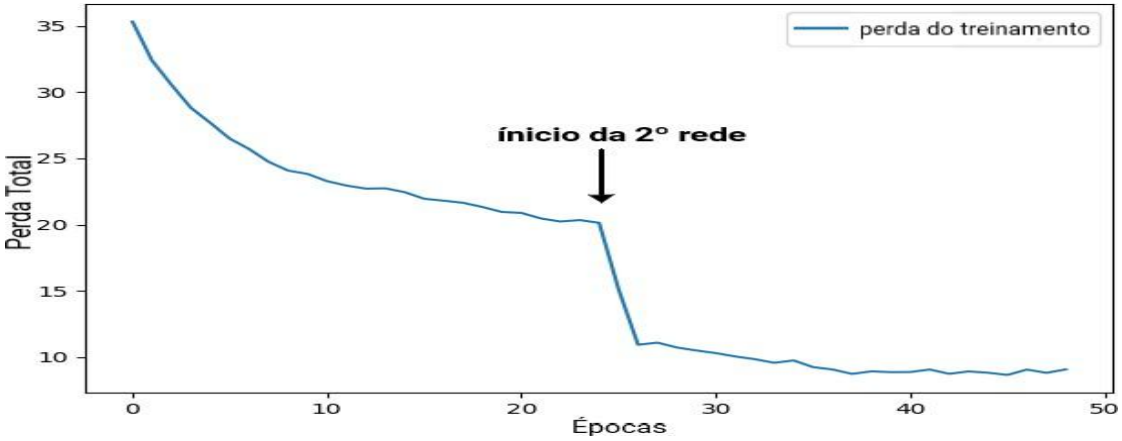
Fonte: adaptada de Carion *et al*. (2020).

Inicialmente, utilizaram-se 35 imagens no treinamento de duas redes buscando a melhor configuração dos hiperparâmetros e para que se constatasse a colaboração entre as redes MLP, assim que a segunda rede fosse inserida. Embora se saiba que para a rede DETR alcançar bons resultados, é necessária uma base significativa de imagens, essa quantidade de amostras foi suficiente para teste da colaboração entres as redes MLPs.

A Figura 17 apresenta um gráfico no qual mostra essa fase, onde se trabalhou com um número pequeno de imagens e também poucas épocas com o objetivo de identificar a colaboração da segunda rede no momento de sua inserção no

treinamento. Pode-se afirmar que a queda considerada um pouco significativa, apresentada no gráfico, acontece em virtude da primeira rede ainda não ter estabilizado o aprendizado e também por se trabalhar com poucas imagens.

Figura 17 - Inserção da rede adicional.



Fonte: Autor.

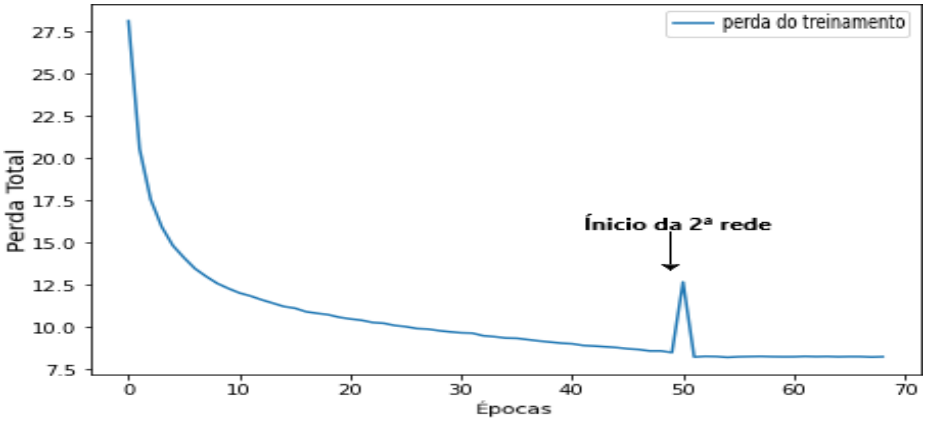
O objetivo do uso de redes adicionais visa a aprender o que a rede principal desconsiderou (DE ALMEIDA, GÓES e NASCIMENTO, 2010). Logo o modelo pode ser visto como uma rede DETR com múltiplas MLPs.

3.3 TREINAMENTO DA DETR COM AS REDES ADICIONAIS

O tamanho do banco de dados influencia diretamente na escolha dos hiperparâmetros. A iniciar pela escolha da taxa de aprendizagem que faz necessário um valor que possibilite a colaboração entre as redes. Levando em consideração um número grande de imagens durante o treinamento, é utilizado uma taxa pequena, que embora demore aparecer bons resultados com apenas uma rede, no momento emque a segunda rede surgir, os resultados serão consideravelmente melhores.

Na Figura 18, é possível observar o treinamento com duas redes onde se possui uma taxa de aprendizagem um pouco alta. Apesar da primeira rede adquirir bons resultados em poucas épocas, com a inserção de uma nova rede ocorre uma instabilidade o qual é refletida em pico no gráficode erro.

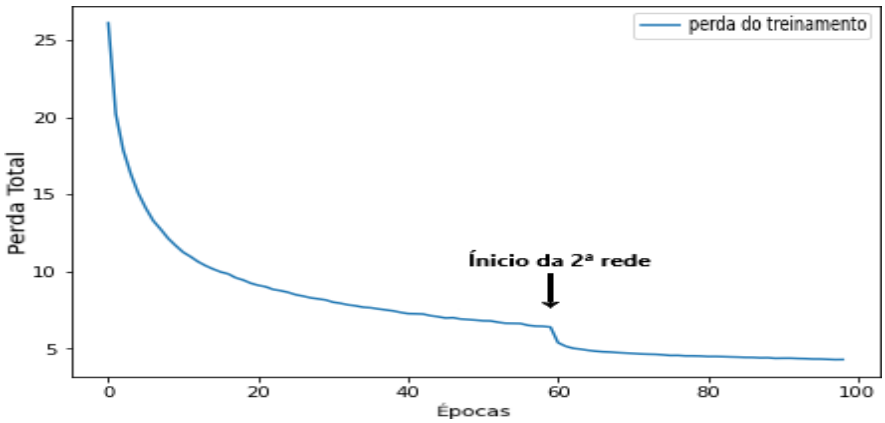
Figura 18 - Efeito de uma taxa de aprendizagem alta com duas redes



Fonte: Autor.

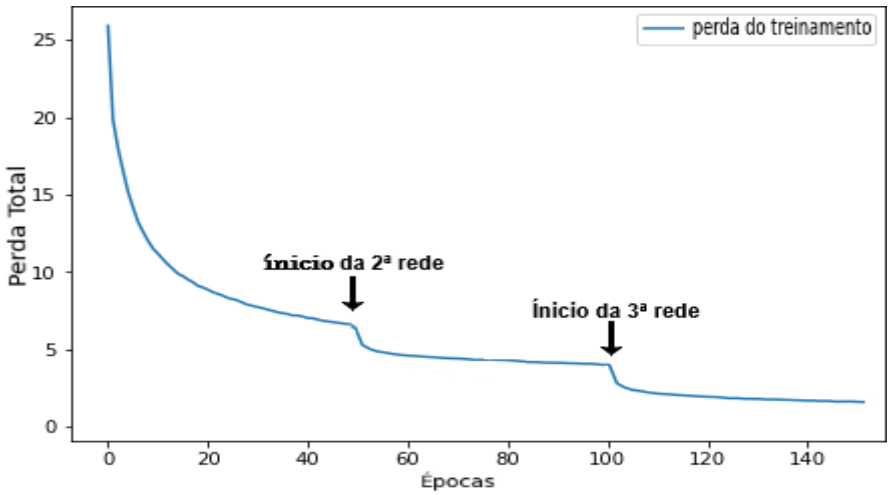
Foram realizadas sessões de treino avaliando diversas configurações da rede buscando, além de bons resultados, uma arquitetura que permitisse um treinamento ágil. Nesse primeiro treinamento com apenas uma rede adicional, colocou-se a rede principal com 60 épocas e a rede adicional com o restante do treinamento. Em cada época, foram utilizadas todas as imagens do conjunto de treino. Após cada 16 épocas, o estado atual da rede é validado com a base de validação e os resultados são salvos. Depois que finalizado o treinamento, realizava-se o teste do desempenho com a base de teste, cuja separação e quantidade foi detalhada previamente. O treino de uma sessão, somando as duas redes, durava cerca de 32 horas. As Figuras 19 e 20 apresentam o treinamento, onde é possível observar a colaboração das redes adicionais.

Figura 19 - Treinamento com apenas uma rede adicional



Fonte: Autor.

Figura 20 - Treinamento com duas redes adicionais



Fonte: Autor.

Depois do treinamento da rede adicional, o modelo completo das redes foi avaliado por meio do conjunto de teste, em que a saída do modelo é a somatória da saída de todas as redes adicionais e a rede principal. Se a primeira rede adicional não possuir resultados significantes, então é inserida uma nova rede e o modelo passa por um novo treinamento. Somente no caso em que a inserção de novas redes adicionais não tiver mais diferenças consideráveis, o treinamento é encerrado.

3.4 AVALIAÇÃO DO CONJUNTO DAS REDES

No presente subcapítulo, são apresentadas as métricas de avaliação usadas neste trabalho, sendo responsável por medir o desempenho da metodologia proposta segundo diferentes critérios.

A avaliação de um sistema de detecção de objetos submete-se uma métrica que leve em consideração detecções corretas, falsas detecções e faltas na detecção em todo o conjunto de teste. A *mean average precision* (mAP) é uma métrica desse tipo que foi introduzida nesse contexto por meio do desafio Pascal VOC (EVERINGHAM *et al*., 2010). O seu cálculo é feito fazendo uso das medidas de precisão e revocação, adquiridos ao aplicar um sistema de detecção em imagens de teste que possuemseus objetos anotados.

Levando-se em consideração que cada imagem contém ou não uma arma de fogo, pode-se julgar o problema de detecção de armas como uma classificação binária. Em virtude disso, as métricas utilizadas neste trabalho são precisão (Pr), sensibilidade também conhecida como *recall* (Se) e F1 Score(F1).

A precisão indica percentualmente quanto o sistema consegue acertar os casos positivos em comparação com os totais positivos apontados pelo sistema; a sensibilidade indica o percentual de acertos dos casos positivos em comparação com todos os casos verdadeiramente positivos; e por fim, a medida F1 Score é uma medida harmônica entre a sensibilidade e a precisão.

Todas as métricas citadas são mostradas nas Equações 2 a 4, onde: VP são os verdadeiros positivos (número de imagens com armas de fogo o qual a rede detectou corretamente pelo menos uma arma de fogo), VN são os verdadeiros negativos (número de imagens que não possuem armas nas quais a rede corretamente não detectou armas de fogo), FP são os falsos positivos (números de imagens que não possuem armas que a rede erroneamente detectou pelo menos uma arma de fogo), e FN são os falsos negativos (número de imagens que contém armas que o modelo erradamente não detectou armas de fogo).

$$Pr = \frac{VP}{VP+FP} \times 100\% \qquad (2)$$

$$Se = \frac{VP}{VP+FN} \times 100\% \qquad (3)$$

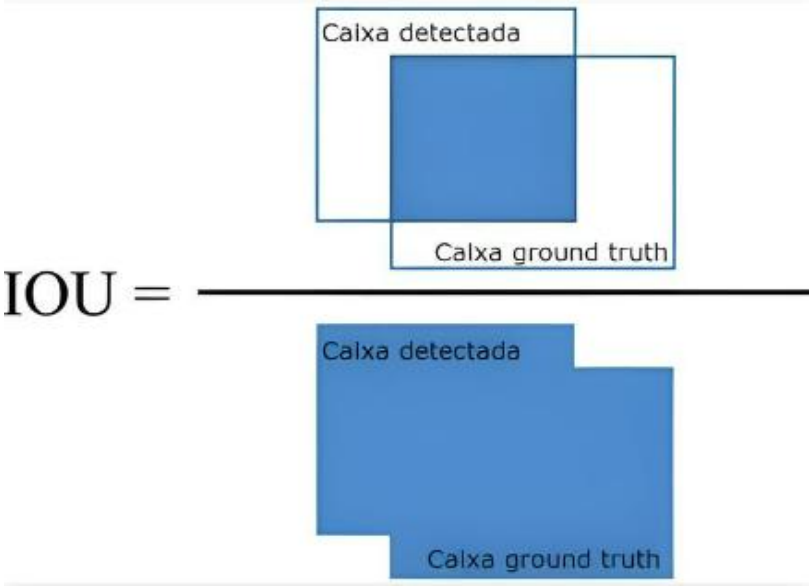
$$F1 = 2 \times \frac{Pr \times Se}{Pr+Se} \qquad (4)$$

O mais adequado é que a precisão e o *recall* sempre possuam valores aproximados, aumentando o grau de confiança do modelo (JENSEN *et al*., 2016), visto que o modelo de predição tem mais de uma classe, o resultado final é a média ponderada do resultado obtido por cada classe.

Levando-se em consideração que a tarefa de deteção de objetos não é uma task de classificação típica, a precisão e o *recall* têm de ser calculados levando em conta um limiar de IoU, ou *intersect over union*, específico. Isto é, para que uma predição possa ser vista como um VP de fato, é preciso que o IoU entre a predição e o objeto seja maior que o limiar escolhido. O limiar determinado para o teste foi 50 %. Sabe-se que o IoU é medido segundo a razão entre a área de interseção do *bounding box* predito e o *ground truth*, e a união das duas *bounding boxes*, conforme

é possível observar no diagrama abaixo:

Figura 21 - Intersecção sobre união



Fonte: Rosebrock (2016).

4 RESULTADOS E DISCUSSÃO

Neste capítulo, são apresentados os resultados adquiridos por meio da metodologia proposta para a tarefa de detecção de armas de fogo em imagens. A eficácia desses modelos é avaliada com base em análises de precisão, *recall* e F1 score, além de uma comparação visual dos resultados obtidos.

Na Tabela 2, é possível observar os melhores resultados obtidos tanto da rede principal, quanto da DETR com uma e com duas redes adicionais. Na Tabela 3, mostram-se os melhores resultados com até quatro redes adicionais. Os valores em cada célula da tabela foram obtidos durante a avaliação com um conjunto de 800 imagens.

Tabela 2 - Resultados 01 de precisão, recall e medida F1

Modelo	Precisão	Recall	F1 score
DETR com 1 MLP	60 %	62 %	0,61
DETR com 2 MLP	69 %	68 %	0,68
DETR com 3 MLP	72 %	96 %	0,82

Fonte: Autor.

Tabela 3 - Resultados 02 de precisão, recall e medida F1

Modelo	Precisão	Recall	F1 score
DETR com 1 MLP	79 %	75 %	0,76
DETR com 2 MLP	81 %	86 %	0,83
DETR com 3 MLP	83 %	94 %	0,88
DETR com 4 MLP	84 %	99 %	0,90

Fonte: Autor.

Conforme mostra a Tabela 2, os valores de precisão variam de 60 % a 83 %. A precisão é uma métrica que mede a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias classificadas como positivas. Nesse caso, os valores de precisão são consistentes, com a maioria dos resultados indicando um desempenho bom do modelo em relação à classificação correta.

Na Tabela 4, é possível observar a matriz de confusão. Os quadrantes de cor verde representam as detecções corretas, os de cor vermelha referem-se às detecções incorretas e os quadrantes de cor cinza, o valor das métricas em percentual. A lacuna gerada é preenchida pelo valor correspondente da matriz de confusão e por uma das cores, em que sua intensidade obedece a ordem da quantidade de redes. Assim sendo, a tonalidade mais clara representa a DETR com uma rede MLP e a tonalidade maisforte com quatro redes MLP autocoordenadas.

Dessa forma, é possível observar a evolução dos valores calculados. Nos quadros de cor verde, mostraram-se melhorias nas detecções corretas no decorrer das inserções das múltiplas redes. O comportamento das detecções incorretas, que estão de cor vermelha, reflete uma queda na medida que é inserida uma nova MLP.

Na diagonal principal, mostra-se o crescimento no valor da acurácia, na qual atingiu o valor 63,2 % com apenas uma MLP e chegou a 83,6 % com as múltiplas redes autocoordenadas. O mesmo ocorre com as métricas de especificidade e recall observadas na horizontal e as de precisão e valor preditivo negativo observadas na vertical.

Com relação ao melhor resultado, das 800 imagens, 669 imagens foram classificadas corretamente onde 663 são verdadeiros positivos e 6 verdadeiros negativos. Das imagens classificadas incorretamente, houve 5 falsos negativos e 126 falsos positivos. As falhas na detecção ocorreram em imagens onde a resolução é de baixa qualidade e luminosidade, entre os principais problemas que dificultam a identificação de armas nos cenários, afetando, diretamente, no desempenho do modelo.

Em um cenário real, o número de falsos positivos não possui muito impacto tanto quanto o número de falsos negativos. É importante considerar o contexto específico em que o modelo de detecção de armas está sendo utilizado. Em certos cenários de alta segurança, como aeroportos ou instalações militares é desejável não ter casos falsos, mas uma vez existentes, é preferível ter um número maior de falsos positivos do que falsos negativos, pois é mais prudente investigar um maior número de casos suspeitos para garantir a segurança. Nesses casos, a eficácia do modelo pode ser medida não apenas pela taxa de falsos positivos, mas também pela taxa de detecção de armas reais. É necessário encontrar um equilíbrio entre minimizar os falsos positivos e garantir que as armas sejam identificadas de forma eficiente, levando em consideração o contexto e as necessidades específicas do

sistema de segurança.

Tabela 4 - Matriz de Confusão

Real	Detectado: com arma	Detectado: sem arma	
com arma	VP 505	FN 163	75,6% / 24,4%
com arma	577	91	86,3% / 13,7%
com arma	629	39	94,1% / 5,9%
com arma	663	5	99,2% / 0,8%
sem arma	FP 131	VN 1	99,2% / 0,8%
sem arma	131	1	99,2% / 0,8%
sem arma	124	8	93,9% / 6,1%
sem arma	126	6	95,4% / 4,5%
	79,4% / 20,6%	99,3% / 0,7%	63,2% / 36,8%
	81,5% / 18,5%	98,9% / 1,1%	72,2% / 27,8%
	83,5% / 16,5%	82,9% / 17,1%	79,6% / 20,4%
	84,3% / 15,9%	45,4% / 54,6%	83,6% / 16,4%

Fonte: Autor.

- DETR com uma MLP
- DETR com duas MLPS
- DETR com três MLPs
- DETR com quatro MLPs

A seguir, busca-se evidenciar, em duas subseções, não apenas a capacidade de identificação correta de armas em diferentes cenários e condições de iluminação, mas também a robustez das redes adicionais em melhorar a detecção da rede principal. O primeiro subcapítulo apresenta os resultados apenas da rede principal, a DETR. Logo após, o segundo discute os resultados da rede principal com as redes adicionais.

4.1 REDE PRINCIPAL

A rede DETR possui uma FFN a qual se resume a uma MLP voltada para localização do *bound box* e para a classificação. Durante a primeira fase, buscou-se manter a configuração original da DETR realizando alterações para adaptação no problema específico deste trabalho que é a detecção de armas de fogos. Após tais configurações, na Figura 22, é possível observar os resultados do teste com os pesos do melhor treinamento. Este treinamento foi realizado por 100 épocas, com taxa de aprendizagem igual a 10^{e-5} e com número de consultas de objetos 14 no qual é o valor que corresponde ao maior número de armas em uma imagem. As arquiteturas

do transformador e da MLP permaneceram iguais à estrutura padrão da DETR.

Figura 22 - Resultados do teste com a rede principal em imagens não canônicas de armas CFTV

Fonte: Autor.

Nesse primeiro teste observa-se que os pesos reconhecem armas de fogo principalmente em imagens onde a arma de fogo está explicitamente exposta. Trazendo para o contexto das imagens de câmeras CFTV, onde o portador da arma a deixa quase oculta e seu tamanho é muito pequeno, o desempenho da rede mostra um nível baixo de confiança, deixando claro espaços para melhorias. Na Figura 23, mostra um exemplo da rede em imagens de câmera CFTV, onde a rede embora reconheça apresença de armas de fogo, mas sua detecção não é correta.

Figura 23 - Resultados do teste com a rede principal em imagens canônicas de armas CFTV

Fonte: Autor.

A detecção incorreta de armas de fogo em imagens de CFTV pela rede principal é um desafio significativo, especialmente quando se trata de objetos

pequenos. Em cenários onde as armas são parcialmente ocultas ou aparecem em ângulos desfavoráveis, a rede pode falhar em identificá-las corretamente, resultando em falsos negativos.

4.2 REDES ADICIONAIS

Como esperado, a rede DETR apresenta um comportamento melhor nas imagens onde a arma de fogo apresenta um tamanho grande. Nas imagens de câmeras CFTV, onde a arma é quase oculta, a rede mostra baixa confiança na detecção. Buscando melhorar o desempenho, em vez de procurar melhorar a configuração dos hiperparâmetros para resolver tal questão, deu-se início à inserção dasredes adicionais.

Neste trabalho, a quantidade usada de redes adicionais serviu para comprovar que o desempenho da detecção melhora significativamente com a inserção de uma nova RNA. Os resultados indicam que o modelo é capaz de detectar com mais precisão a maioria das armas de fogo presentes nas imagens e com uma taxa de falsos positivos relativamente baixa. Além do mais, cada inserção de uma RNA melhorou a capacidade de localizar objetos mesmo quando ocorre a oclusão parcial da arma, como é possível observar na Figura 24.

Figura 24 - Comparação da detecção da rede DETR com as redes adicionais com oclusão parcial da arma

Fonte: Autor.

Diante dos testes realizados, embora também não se tenha utilizado elevado número de épocas durante o treinamento, fator este importante para o modelo convergir, é possível notar que o desempenho do modelo melhora com a inserção de uma rede adicional, porém é mais nítida a melhoria a partir do acréscimo da segunda

rede adicional. Na Figura 25, é possível observar uma comparação do modelo com a inserção das redes.

Figura 25 - Comparação da detecção da rede DETR com as redes adicionais.

Fonte: Autor.

Na primeira detecção com apenas a rede DETR, o *bound box* ocupa a maior parte da imagem, não localizando corretamente a arma de fogo na imagem. Quando inserida uma rede adicional, é possível notar que, embora a localização da detecção não tenha sido correta, houve uma diminuição no tamanho do *bound box* e uma aproximação da localização correta do *bound box*. Com duas redes adicionais, houve significativa melhora no comportamento da detecção mostrando a localização corretada arma presente e com maior percentual de precisão.

Na figura abaixo, são mostrados alguns outros casos de detecções com as redes adicionais em ambientes diferentes em imagens de CFTV. É importante frisar que essas imagens possuem uma boa qualidade e luminosidade adequada.

Figura 26 - Detecção em outros cenários com as redes adicionais.

Fonte: Autor.

As falhas de detecção ocorreram em situações em que a imagem é de baixa qualidade ou onde o ambiente é escuro. Também houve falhas com objetos que geram falsos positivos como celulares nas mãos em imagens de baixa resolução. Apesar de que alguns casos de erro tenham aparecido, esses são mínimos e não inviabilizam a colaboração entre as redes.

4.3 COMPARAÇÃO COM A LITERATURA

Neste subcapítulo, comparam-se os resultados obtidos com outros trabalhos encontrados na literatura. A Tabela 5 mostra essa comparação, apresentando valores das métricas obtidas. É importante salientar que este trabalho não desejou superar o estado da arte do problema de detecção de armas de fogo em imagens, mas sim apresentar uma pesquisa na qual pode contribuir com a evolução dos detectores de objetos em imagens, podendo ser aplicado em outros modelos de detecção.

Além disso, vale enfatizar que cada trabalho aborda diferentes contextos, utilizam bases de dados distintas e metodologias variadas, o que torna inviável uma conclusão direta sobre qual é o mais eficiente.

Ao comparar o modelo deste trabalho com outros descritos na literatura, embora essa comparação não seja bem nivelada, essa análise fornece um quadro útil para entender como este trabalho se posiciona em relação aos métodos similares já desenvolvidos. O intuito principal dessa comparação é destacar as diversas abordagens e técnicas utilizadas na detecção de armas de fogo em imagens, proporcionando uma análise crítica das vantagens e desafios enfrentados por cada método. Reconhece-se que as variações nas condições experimentais, nas bases de dados e nos cenários de aplicação influenciam significativamente os resultados obtidos, enfatizando a importância de uma avaliação cuidadosa e contextualizada ao interpretar os estudos existentes nesta área. Os trabalhos que não incluem valores na métrica F1 Score indicam que os autores não a utilizaram em suas pesquisas.

O método proposto para detecção de armas apresentou resultados promissores quando comparado a outros trabalhos com resultados inferiores. Com uma precisão de 84 % e recall de 99 %, o método destaca-se em termos de acurácia na detecção de armas. Além disso, o F1 Score de 0,909 indica um equilíbrio entre precisão e recall, demonstrando a eficiência do método em identificar corretamente

armas em diferentes cenários.

Ao comparar com estudos anteriores, como o trabalho de Baldessar (2021), que obteve uma precisão de 96,5 % e recall de 84,5 %, ou o estudo de Ashraf *et al*. (2021), com precisão de 81 % e recall de 99 %, o método proposto apresenta resultados competitivos.

É importante ressaltar que cada método de detecção de armas possui suas próprias características e limitações e a escolha do mais adequado dependerá do contexto específico de aplicação. No entanto, os resultados obtidos pelo método proposto indicam seu potencial como uma solução confiável e eficiente para a detecção de armas.

O estudo de Ruprah *et al*. (2023) avaliou três métodos de detecção de armas: YOLO, SSD e Faster R-CNN. Os resultados mostraram que o Faster R-CNN obteve uma precisão de 92 % e recall de 95 %, enquanto o YOLO apresentou uma precisão de 86 % e recall de 90 %. Em comparação, o método proposto alcançou uma precisão de 83 % e recall de 94 % superando os resultados do modelo SSD.

Além disso, é importante destacar que cada método pode ter suas próprias vantagens e desvantagens em termos de eficiência computacional, complexidade de implementação e requisitos de recursos. Portanto, a escolha entre o método proposto e o Faster R-CNN dependerá das necessidades específicas do projeto e dos critérios de avaliação estabelecidos. No geral, tanto o método proposto, quanto o Faster R-CNN, avaliado por Ruprah *et al*. (2023), são abordagens promissoras para a detecção de armas, mostrando bons resultados.

O estudo realizado por Ashraf *et al*. (2021) reportou uma precisão de 81 % e recall de 99 % em sua metodologia de detecção de armas. Em contraste, o método proposto obteve uma precisão superior, atingindo 83 %, com uma diferença de 2 %. Isso sugere que o método proposto pode ser mais preciso na identificação de armas, evitando falsos positivos.

Tabela 5 - Comparação com a literatura

Método	**Precisão**	**Recall**	**F1 score**
Ruprah *et al*. (2023)	YOLO 86% SSD 82% Faster R-CNN 92%	YOLO 90% SSD 86% Faster R-CNN 95%	YOLO 0,88 SSD 0,84 Faster R-CNN0,93
Baldessar (2021)	96,5%	84,5%	-
Qi *et al*. (2021)	95,99%	99,69%	-
Olmos *et al.* (2018)	84,21%	100%	-
Ruiz-Santaquiteria *et al.* (2021)	91,88%	83,83%	-
González *et al*. (2020)	88,12%	100%	0,93
Olmos *et al*. (2021)	100%	80%	0,89
Ashraf *et al.* (2021)	81%	99%	0,89
Al-Mousa *et al.* (2023)	93,4%	91,5%	0,92
Kambhatla *et al*. (2022)	89,4%	70,1%	-
Método Proposto	**84%**	**99%**	**0,91**

Fonte: Autor.

4.3.1 Comparação com trabalhos que fizeram uso do banco de dados Monash Guns

Na comparação a seguir, todos os métodos foram testados usando o conjunto de dados Monash Guns extraídos do trabalho Ruiz-Santaquiteria *et al*. (2021), conforme apresentado na Tabela 6. Na base de teste dos trabalhos mencionados, foram utilizadas 300 imagens, enquanto no método proposto foram utilizadas 800. Essa diferença no número de imagens utilizadas nos testes entre o método proposto e os trabalhos mencionados pode influenciar a interpretação dos resultados. No entanto, é importante ressaltar que, ao interpretar os resultados, deve-se considerar não apenas a quantidade de dados, mas também a representatividade e diversidade da base de teste para garantir uma análise justa e precisa do desempenho dos métodos comparados.

O método proposto demonstra um equilíbrio notável entre precisão e sensibilidade em comparação com os outros trabalhos analisados. Embora tenha uma precisão menor em comparação com os resultados de Ruiz-Santaquiteria *et al*. (2021), sua sensibilidade é significativamente superior, atingindo 99 %. Isso sugere

que o método proposto tem a capacidade de identificar uma vasta maioria dos casos relevantes, minimizando falsos negativos, o que pode ser crucial em aplicações onde a detecção de padrões específicos é essencial.

Tabela 6 - Comparação com trabalhos que fizeram uso do banco de dados Monash Guns

Método	**Precisão**	**Recall**
Ruiz-Santaquiteria *et al.* (2021)	96,83%	33,67%
Velasco-Mata *et al*. (2021)	78,48%	20,67%
Basit *et al*. (2020)	89,29%	8,33%
Salido *et al*. (2021)	88,33%	17,67%
Método Proposto	**84%**	**99%**

Fonte: Autor.

Em outros termos, ao apresentar uma maior taxa de falso positivo em comparação com o trabalho de Ruiz-Santaquiteria *et al*. (2021), o método proposto oferece vantagens significativas em cenários de monitoramento de segurança, especialmente na detecção de armas de fogo em imagens. Essa abordagem, embora menos precisa, possui uma sensibilidade significativa, permitindo que operadores de CFTV identifiquem rapidamente potenciais ameaças. Por exemplo, em um ambiente de grande circulação, como um *shopping center*, um alerta gerado por esse método pode sinalizar a presença de uma arma, permitindo que a equipe de segurança intervenha antes que a situação se agrave. Mesmo que esse alerta possa ser um falso positivo, a prioridade de agir rapidamente pode salvar vidas.

A desconsideração de um alerta proveniente do método proposto pode ter consequências graves. Em um cenário, onde um operador de CFTV ignorando um sinal de detecção de arma durante um evento público, como um *show*. Se o alerta se referir a uma situação real, a inação pode resultar em um incidente violento, colocando em risco a segurança de centenas de pessoas. Assim, a maior taxa de falso positivo do método proposto se justifica em contextos onde a detecção de armas é crítica, já que priorizar a sensibilidade pode ser fundamental para a proteção da comunidade e a prevenção de tragédias.

Em contraste com Velasco-Mata *et al*. (2021), o método proposto apresenta uma melhoria em ambas as métricas. Enquanto Basit *et al*. (2020) apresenta uma precisão de 89,29 %, o método proposto alcança 84 % de precisão. Essa taxa de precisão demonstra a eficácia do método proposto na identificação precisa de padrões relevantes no conjunto de dados, mesmo que sua precisão seja inferior à de Basit *et al*. (2020). Essa combinação de alta sensibilidade e precisão competitiva ressalta o valor do método proposto em aplicações onde a detecção de padrões específicos é crucial.

Embora a precisão do método proposto seja inferior à de Salido *et al.* (2020), que atinge 88,33 %, a sensibilidade muito superior do método proposto indica uma capacidade superior na identificação abrangente de casos positivos. Essa comparação sugere que o método proposto pode ser particularmente valioso em cenários onde a minimização de falsos negativos é crucial, superando os métodos anteriores ao equilibrar de maneira eficaz a precisão e a sensibilidade.

5 CONCLUSÃO

Este trabalho apresentou a viabilidade da utilização da técnica múltiplas RNAs autocoordenadas com *Detection Transformer* para criação de um detector genérico de armas de fogo. O método utilizado mostrou-se robusto, sendo capaz de detectar corretamente armas que não foram apresentadas durante o treinamento, sendo estas de modelos mais comuns de armas e em diversos ambientes. Além do mais, apresentou melhor performance em imagens de CFTV, diferente da rede principal.

O modelo resultante deste projeto tem a capacidade de facilitar ainda mais o processo de detecção de armas. Graças aos avanços na visão computacional, existem muitos benefícios nesta tecnologia. Por exemplo, o modelo apresentado neste trabalho possui um aprendizado acumulativo, o qual reduziu significativamente o tempo de treinamento e ainda entregando uma simplicidade em termos de carga computacional. Além disso, oferece uma significativa redução no tempo de projeto, encurtando o período desde o início do desenvolvimento da solução até a descoberta da solução final.

Como consequência dessa redução, esse avanço permite explorar diferentes configurações de rede de forma iterativa, não necessitando encontrar uma configuração ideal única. A primeira rede pode ser implementada com uma configuração inicial menos otimizada, enquanto as redes subsequentes são capazes de refiná-la progressivamente, melhorando continuamente o desempenho do sistema. Essa abordagem iterativa não só acelera o processo de desenvolvimento, mas também resulta em soluções mais robustas e eficientes, adaptadas às necessidades específicas do problema em questão.

Entretanto esse método apresenta pontos ainda a serem melhorados na detecção de objetos, devido à inicialização dos pesos na qual uma parte desses pesos tem que ser zerados. Isso pode retardar o desempenho até que as redes subsequentes possam melhorar o modelo através da otimização iterativa. Além disso, o uso de muitas redes nessa abordagem pode aumentar o risco de *overfitting*. Portanto, é crucial equilibrar o número de redes para garantir que as melhorias incrementais traduzam-se em ganhos reais de desempenho sem comprometer a capacidade do modelo generalizar.

Embora não se tenham utilizado muitas redes adicionais ou muitas épocas nos treinamentos para análise do comportamento do modelo, os resultados

encontrados foram bastante satisfatórios para a detecção de armas de fogo. Em resumo, uma nova rede focou em aprender os conhecimentos ignorados pelas redes anteriores durante o treinamento. Assim, o modelo alcançou 84 % de precisão na detecção de armas, o que comprovou a eficiência da técnica utilizada.

5.1 CONTRIBUIÇÕES DESTE TRABALHO

A técnica utilizada neste trabalho apresentou melhorias significativas na detecção de armas em imagens CFTV, fator este no qual é um problema conhecido ao se estudar as dificuldades da rede DETR, comprovando que a técnica tem aplicabilidade. Além disso, contribuiu diretamente com as pesquisas na área de segurança pública relacionadas ao desenvolvimento de ferramentas computacionais para detecção de situações potencialmente perigosas auxiliando no combate à violência.

O trabalho em questão oferece contribuições significativas em várias frentes. Primeiramente, ao empregar a técnica de múltiplas redes autocoordenadas com DETR, há uma notável redução no tempo de projeto, agilizando desde o desenvolvimento inicial até a obtenção da solução final. Além disso, destaca-se pela melhoria do desempenho na detecção de objetos pequenos, como armas de fogo em imagens, onde precisão e rapidez são essenciais para a segurança pública. Essas melhorias não apenas fortalecem a capacidade de resposta em situações críticas, mas também colaboram diretamente para aprimorar a eficácia das medidas de segurança, proporcionando um ambiente mais seguro e protegido para a sociedade como um todo.

5.2 TRABALHOS FUTUROS

Essa pesquisa apresentou resultados bastante interessantes para o contexto de detecções de armas de fogo fazendo uso da técnica de MRNA autocoordenadas com a rede DETR. No entanto, diversas novas possibilidades de investigação e de melhoria podem ser apontadas a partir do que foi feito, como a aplicação dessa mesma técnica com foco na colaboração entre transformadores. Outro fator interessante também seria, de forma similar, aprofundar mais com relação ao aumento no número de camadas da MLP. Esses seriam alguns pontos interessantes para análise de melhorias com a rede *Detection Transformer*.

Outra possibilidade seria utilizar a técnica de MRNA autocoordenadas em outros contextos, seja na área da medicina ou agricultura e aplicando com outras redes nas quais venham se destacando no campo da visão computacional. Enfim, este trabalho mostrou uma técnica de aprendizado que até então não havia sido aplicada com redes que fazem uso do transformador para detecção de objetos, abrindo um vasto campo de pesquisas e incentivando o desenvolvimento de tecnologias inovadoras.

REFERÊNCIAS BIBLIOGRÁFICAS

ASHRAF, A. H. *et al*. Weapons Detection for Security and Video Surveillance Using CNN and YOLO-V5s. **Computers, Materials and Continua**, Henderson, v. 70, n. 2, p. 2761-2775, 2021. Disponível em: http://doi.org/10.32604/cmc.2022.018785. Acesso em: 15 jan. 2022.

AKDOGAN, A. **DETR**: End-to-End Object Detection with Transformers and Implementation of Python. Disponível em: http://towardsdatascience.com/detr-end-to-end-object-detection-with-transformers-and-implementation-of-python-8f195015c94d/. Acesso em: 12 out. 2022.

AL-MOUSA, A. *et al*. Deep Learning-Based Real-Time Weapon Detection System. **International Journal of Computing and Digital Systems**, Bahrain, v. 20, p. 2210-142, 2023. Disponível em: http://dx.doi.org/10.12785/ijcds/14014. Acesso em: 10 jan. 2022.

BALDESSAR, Gabriel. **Detecção de Armas de Fogo em Vídeo através de Redes Neurais**. Disponível em: http://repositorio.ufsc.br/handle/123456789/223679. Acesso em: 14 set. 2021.

BASIT, A. *et al.* **Localizing firearm carriers by identifying human-object pairs**. Disponível em: http://arxiv.org/abs/2005.09329. Acesso em: 16 set. 2023.

BOESCH, G. **Object Detection in 2024**: The Definitive Guide. Disponível em: http://viso.ai/deep-learning/object-detection/. Acesso em: 15 fev. 2024.

CARION, N. *et al.* End-to-end object detection with transformers. *In:* EUROPEAN CONFERENCE ON COMPUTER VISION, 2020, Glasgow: **Proceedings [...]**. Glasgow: ACM, 2020. p. 213–229.

CHATTERJEE, R. *et al*. A Deep Learning-based Efficient Firearms Monitoring Technique for Building Secure Smart Cities. **IEEE Access: Practical Innovations**, **Open Solutions**, Australia, v.11, p. 37515-37524, 2023. Disponível em: http://ieeexplore.ieee.org/document/10099460. Acesso em: 11 fev. 2022.

CHO, Sung-Bae, *et al*. Combining multiple neural networks by fuzzy integral for robust classification. **IEEE Transactions on Systems, Man and Cybemetrics**, Memphis, v. 25, n. 2, p. 380-384, 1995. Disponível em: http://ieeexplore.ieee.org/document/364825. Acesso em: 5 jun. 2022.

CHO, Sung-Bae. Fusion of neural networks with fuzzy logic and genetic algorithm. **Integrated Computer Aided Engineering**, Califórnia, v. 9, n. 4, p.363-372, 2002. Disponível em: http://content.iospress.com/articles/integrated-computer-aided-engineering/ica00128. Acesso em: 5 jun. 2022.

DAI, J. *et al*. R-FCN: object detection via region-based fully convolutional networks. *In*: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2016, New York. **Proceedings[...].** New York: ACM, 2016. p. 379- 387.

DANG, L. M. *et al.* DefectTR: End-to-end defect detection for sewage networks using a transformer. **Construction and Building Materials**, Valencia, v. 325, n.126584, p.126584, 2022. Disponível em: http://doi.org/10.1016/j.conbuildmat.2022.126584. Acesso em: 21 jun. 2022.

DE ALMEIDA NETO, A. **Aplicações de Múltiplas Redes Neurais em Sistemas Mecatrônicos**. 2003. 154 f. Doutorado (Curso de Engenharia Aeronáutica e Mecânica) - Instituto Tecnológico de Aeronáutica, São José dos Campos, 2003.

DE ALMEIDA NETO, A. *et al*. Accumulative learning using multiple ANN for flexible link control. **IEEE Transactions on Aerospace and Electronic Systems**, Bedfordshire, v. 46, n. 2, p. 508-524, 2010. ISSN 00189251.

DEEPTHI, T. *et al.* Firearm recognition using convolutional neural network. **International Journal of Scientific Research in Computer Science**, **Engineering and Information Technology**, India, p.136-141, 2019. Disponível em: http://ijsrcseit.com/paper/CSEIT 195226.pdf. Acesso em: 15 abr. 2022.

EGIAZAROV, A. *et al*. Firearm detection and segmentation using an ensemble of semantic neural networks. *In*: EUROPEAN INTELLIGENCE AND SECURITY INFORMATICS CONFERENCE, 19., 2020, Norway, **Proceedings [...]**. Norway: IEEE, 2020. p. 70-77**.**

EVERINGHAM, M. *et al*. The Pascal Visual Object Classes (VOC) Challenge. **International Journal of Computer Vision**, Massachusetts, v. 88, n. 2, p. 303–338, 9 Set. 2009. Disponível em: http://doi.org/10.1007/s11263-009-0275-4. Acesso em: 15 jan. 2022.

GAO, M. *et al.* A Novel Deep Convolutional Neural Network Based on ResNet-18 and Transfer Learning for Detection of Wood Knot Defects. **Journal of sensors**, Malásia, v. 2021, n.1, 2021. Disponível em: http://doi.org/10.1155/2021/4428964. Acesso em: 21 jan. 2022.

GIRSHICK, R. *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. *In*: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 14., Columbus, **Proceedings [...]**. Columbus: IEEE, 2014. p. 580-587.

GONZÁLEZ, J. L. S. *et al.* Real-time gun detection in CCTV: An open problem. **Neural Networks**, Ohio, v. 132, p. 297-308, Dec. 2020. Disponível em: http://doi.org/10.1016/j.neunet.2020.09.013. Acesso em: 10 ago. 2022.

GUAN, T. *et al*. M3detr: Multirepresentation, multi-scale, mutual-relation 3d object detection with transformers. *In*: IEEE/CVF WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION, 241., Waikoloa: IEEE, 2021. p. 2293–2303.

HASANAH, SA *et al*. A Deep Learning Review of ResNet Architecture for Lung Disease Identification in CXR Image. **Applied sciences**, Basileia, v. 13, n. 24, p. 13111, 2023. Disponível em: http://doi.org/10.3390/app132413111. Acesso em: 10 ago. 2022.

HASHEM, S. *et al.* Optimal Linear Combination of Neural Networks: An Overview. *In*: IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE, 94., Orlando: IEEE, 1994. v. 3, p.1507-1512.

HASHEM, S.; SCHMEISER, B., Improving Model Accuracy Using Optimal Linear Combinations of Trained Neural Networks. **IEEE Transactions on Neural Networks**, v. 6, n. 3, p. 792-794, 1995. Disponível em: http://doi.org/10.1109/72.377990. Acesso em: 10 ago. 2022.

HASHEM, S. Algorithms for Optimal Linear Combinations of Neural Networks. *In*: INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, 1997, Houston, **Proceedings [...]**. Houston: IEEE, 1997. v. 1, p. 242–247.

HE, L. *et al*. End-to-end video object detection with spatial-temporal transformers. MM '21. *In*: INTERNATIONAL CONFERENCE ON MULTIMIDEA, 29., 2021, New York, **Proceedings[...]**. New York: ACM, 2021. p.1507- 1516.

HUDA, M. **Paper Review**: DETR (Detection Transformer). Disponível em: http://iniakunhuda.com/posts/paper-review-detr/. Acesso em: 10 jun. 2023.

JENSEN, M. B. *et al*. Vision for Looking at Trafic Lights: Issues, Survey, and Perspectives. **IEEE Transactions on Intelligent Transportation Systems**, Genova. v.17, n. 7, p. 1800-1815, 2016. ISSN 1524-9050. Disponível em: http://doi.org/10.1109/TITS.2015.2509509. Acesso em: 5 set. 2022.

KAMBHATLA, A. *et al.* Firearm detection using deep learning. **Intelligent Systems with Applications**, Amsterdam, p. 200-218, 2022. Disponível em: http://doi.org/10.1007/978-3-031-16075-2_13. Acesso em: 10 fev. 2023.

LIM, J. *et al.* Gun Detection in Surveillance Videos using Deep Neural Networks. *In:* ASIA-PACIFIC SIGNAL AND INFORMATION PROCESSING ASSOCIATION ANNUAL SUMMIT AND CONFERENCE, 2019, Lanzhou: **Proceedings [...]**. Lanzhou: IEEE, Nov. 2019. p.18-21.

LIM, J. *et al*. Deep multilevel feature pyramids: Application for non-canonical firearm detection in video surveillance. **Engineering applications of artificial intelligence**, França, v. 97, n.104094, 2021. Disponível em: http://doi.org/10.1016/j.engappai.2020.104094. Acesso em: 1 out. 2022.

LIN, T.-Y. *et al.* Focal Loss for Dense Object Detection. *In*: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2017, Venice: **Proceedings [...]**. Venice: IEEE, 2017. p. 2999-3007.

MISRA, I. *et al*. An end-to-end transformer model for 3D object detection. *In:* INTERNATIONAL CONFERENCE ON COMPUTER VISION, 21., 2021, Montreal: **Proceedings [...]**. Montreal: IEEE, 2021. p. 2906-2917.

NIKOUEI, S. Y. *et al*. Real-Time Human Detection as an Edge Service Enabled by a Lightweight. *In*: IEEE INTERNATIONAL CONFERENCE ON EDGE COMPUTING, 18., 2018, São Francisco: **Proceedings [...]**. São Francisco: IEEE, 2018. p. 125-129.

OLMOS, R. *et al.* Automatic handgun detection alarm in videos using deep learning. **Neurocomputing**, Middlesex, v. 275, p. 66-72, 2018. Disponível em: http://doi.org/10.1016/j.neucom.2017.05.012. Acesso em: 11 out. 2022.

OLMOS, R. *et al*. **MULTICAST**: MULTI Confirmation-level Alarm SysTem based on CNN and LSTM to mitigate false alarms for handgun detection in video-surveillance. Disponível em: http://arxiv.org/abs/2104.11653. Acesso em: 20 out. 2021.

PERRONE, M. P. *et al*. When networks disagree: Ensemble methods for hybrid neural networks. **World Scientific Series in 20th Century Physics**, Hackensack, p. 342- 358, 1995. Disponível em: http://doi.org/10.1142/9789812795885_0025. Acesso em: 12 dez. 2022.

QI, D. *et al*. A dataset and system for real-time gun detection in surveillance video using deep learning. *In*: IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS, 21., p. 667-672, 2021.

REDMON, J. *et al.* You Only Look Once: Unified, Real-Time Object Detection. *In*: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016, Las Vegas: **Proceedings [...]**. Las Vegas: IEEE, 2016. p. 779-788.

RUIZ-SANTAQUITERIA, J. *et al.* Handgun detection using combined human pose and weapon appearance. **IEEE access: practical innovations, open solutions**, Massachusetts, v. 9, p. 123815-123826, 2021. Disponível em: http://doi.org/10.1109/ACCESS.2021.3110335. Acesso em: 2 mai. 2023.

RUSSAKOVSKY, O. *et al*. Imagenet large scale visual recognition challenge. **International Journal of Computer Vision**, v. 115, n. 3, p. 211–252, 2015. Disponível em: http://doi.org/10.1007/s11263-015-0816-y. Acesso em: 15 mai. 2023.

ROSEBROCK, A. **Intersection over Union (IoU) for object detection** - PyImageSearch. 2016. Disponível em: http://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/. Acesso em: 5 jul. 2021.

RUPRAH, T. *et al.* **Crime Prediction Based on Person-Weapons Relation using DeepLearning Techniques**. Disponível em: http//assets-eu.researchsquare.com/files/rs-2743470/v1/efd549af-961b-45cf-b25e-1ca93a128d25.pdf?c=1693777942. Acesso em: 2. jun. 2023.

SALIDO, J. *et al.* Automatic handgun detection with deep learning in video surveillance images. **Applied Sciences**, Basileia, v.11, n.13, p. 6085, Jun. 2021. Disponível em: http://doi.org/10.3390/app11136085. Acesso em: 16. jun. 2022.

SUN, Z. *et al*. Rethinking transformer-based set prediction for object detection. *In*: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 21., 2020, Montreal: **Proceedings [...]**. Montreal: IEEE, 2020. p. 3591-3600.

SHAH, D. **Revolutionary Object Detection Algorithm from Facebook AI**. Disponível em: http://medium.com/visionwizard/detr-b677c7016a47. Acesso em: 11 jul. 2023.

TEIXEIRA, Aline Porfiro. **Múltiplas Redes Neurais Utilizando Redes MLP e RBF**. Disponível em: http://repositorio.uema.br/handle/123456789/464. Acesso em: 11 fev. 2022.

VASWANI, A. *et al*. Attention Is All You Need. *In*: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, California: **Proceedings [...]**. California: ACM, 2017. p. 5998-6008.

VELASCO-MATA, A. *et al*. Using human pose information for handgun detection. **Neural Computing and Applications,** United Kingdom, v. 33, n. 24, p. 17273-17286, 2021. Disponível em: http://doi.org/10.1007/s00521-021-06317-8. Acesso em: 1 jun. 2022.

VERMA, G. K. *et al*. A handheld gun detection using faster r-cnn deep learning. *In*: INTERNATIONAL CONFERENCE ON COMPUTER AND COMMUNICATION TECHNOLOGY, 7., 2017, Allahabad: **Proceedings [...]**. Allahabad: ACM, 2017. p. 84-88.

WAISELFISZ, J. J. **Mapa da violência 2016**: Homicídio por armas de fogo no Brasil. Brasília: Flacso Brasil, 2016.

XU, X. *et al.* 3D human texture estimation from a single image with transformers. *In*: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 21., 2021, Montreal: **Proceedings [...]**. Montreal: IEEE, 2021. p. 13829-13838.

YANG, Y. Y. *et al*. Steel yield strength prediction using ensemble model-performance improvement over single neural network model. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT PROCESSING AND MANUFACTURING OF MATERIALS**,** n.1, Vancouver: **Proceedings [...]**. Vancouver: Springer, p. 335-346, 2001.

ZHOU, Q. *et al.* **TransVOD:** End-to-End Video Object Detection with Spatial-Temporal Transformers. Disponível em: http://arxiv.org/pdf/2201.05047.pdf. Acesso em: 20 jan. 2022.