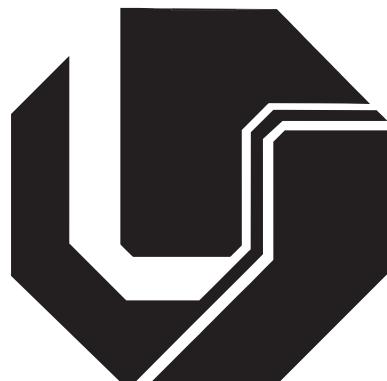


UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA ELÉTRICA
PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA



**Inserção Automática de Componentes
em Ambientes Virtuais de Treinamento para Substações
de Energia utilizando Inteligência Artificial**

Leandro Sena Zuza

Uberlândia

2024

Leandro Sena Zuza

**Inserção Automática de Componentes
em Ambientes Virtuais de Treinamento para Subestações
de Energia utilizando Inteligência Artificial**

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para obtenção do Título de Mestre em Ciências.

Julho X, 2024.

Membros da Banca:

Prof. Alexandre Cardoso, Dr.
Orientador - UFU

Prof. Daniel S. Caetano, Dr.
Coorientador - UFU

Agradecimentos

[INSERIR NOVOS AGRADECIMENTOS]

“Então considerei que as botas apertadas são uma das maiores venturas da Terra,
porque, fazendo doer os pés, dão azo ao prazer de as descalçar.”
(Machado de Assis, em Memórias Póstumas de Brás Cubas)

Abstract

The precise identification of equipment in images plays a crucial role in various operations related to power substations, facilitating not only maintenance but also the monitoring of these facilities. In this work, we present results on the efficiency of YOLOv8 in detecting equipment present in power substations from images obtained by Unmanned Aerial Vehicles (UAVs). We employ different optimization techniques to enhance detection efficiency, aiming to achieve more accurate and faster results. Additionally, this study aims to go beyond mere training with captured photos, seeking to identify the best-trained model to create a script capable of selecting, from a database of virtual reality models, the elements necessary for assembling a virtual power substation. Thus, we aim not only to improve the maintenance and monitoring processes of power substations in physical reality but also to streamline and enhance the generation of Virtual Training Environments for procedures related to these substations. With this advancement, we hope to not only optimize the use of detection technology in power substations but also to significantly contribute to the creation of realistic and efficient virtual environments for training in procedures related to the operation and maintenance of these facilities.

Keywords

Keywords - Power Substation; UAV; YOLOv8; Optimization; Virtual Training Environments

Resumo

A identificação precisa de equipamentos em imagens desempenha um papel crucial em várias operações relacionadas às subestações de energia, facilitando não apenas a manutenção, mas também o monitoramento dessas instalações. Neste trabalho, apresentamos resultados da eficiência da YOLOv8 na detecção de equipamentos presentes em subestações de energia, a partir de imagens obtidas por Veículos Aéreos Não Tripulados (VANTs). Utilizamos diferentes técnicas de otimização para aprimorar a eficiência na detecção, visando alcançar resultados mais precisos e rápidos. Além disso, este estudo visa ir além do mero treinamento com as fotos capturadas, buscando identificar o melhor modelo treinado para criar um script capaz de selecionar, a partir de uma base de modelos de realidade virtual, os elementos necessários para a montagem de uma subestação de energia virtual. Dessa forma, almejamos não apenas melhorar os processos de manutenção e monitoramento das subestações de energia na realidade física, mas também agilizar e aprimorar a geração de Ambientes Virtuais de Treinamento para procedimentos relacionados a essas subestações. Com este avanço, esperamos não só otimizar o uso de tecnologia de detecção em subestações de energia, mas também contribuir significativamente para a criação de ambientes virtuais realistas e eficientes para treinamento em procedimentos relacionados à operação e manutenção dessas instalações.

Palavras Chave

Subestação de Energia; VANTs; YOLOv8; Otimização; Ambientes Virtuais de Treinamento

Publicações

As publicações relacionadas à pesquisa e ao trabalho realizado são listadas a seguir:

1. Zuza, L. S., Cardoso, A., Lamounier, E., & Caetano, D. (2024). Análise dos Parâmetros da YOLOv8 na eficiência da identificação de reatores de núcleo de ar a partir de imagens de subestações de energia. In: *CISTI'2024 - 19ª Conferência Ibérica de Sistemas e Tecnologias de Informação*, 25-28 de junho de 2024, Salamanca, Espanha.

Sumário

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

VANT	Veículo Aéreo Não Tripulado
RV	Realidade Virtual
NA	Neurônio Artificial
RNC	Rede Neural Convolucional
YOLO	You Look Only Once
GD	Gradiente Descendente
SGD	Stochastic Gradient Descent
ADAM	Adaptive Moment Estimation

1 Introdução

1.1 Motivação

As subestações de energia desempenham um papel fundamental no sistema de distribuição de energia elétrica no Brasil, permitindo a transferência eficiente e segura de eletricidade entre diferentes níveis de tensão. Elas são cruciais para garantir que a eletricidade gerada em usinas seja entregue aos consumidores com a qualidade e confiabilidade necessárias. Efetivamente, desempenham um papel crucial na estabilidade do sistema elétrico, facilitando a manutenção, controle e proteção da rede. Seu funcionamento envolve várias etapas, começando com a recepção da eletricidade gerada em usinas de energia. A eletricidade é então transformada em níveis de tensão adequados para distribuição por meio de transformadores. Nas subestações, também ocorrem operações de chaveamento, onde os dispositivos de comutação são usados para controlar o fluxo de eletricidade e direcioná-lo para as áreas desejadas da rede. As subestações também estão equipadas com sistemas de proteção que detectam e isolam falhas para evitar danos ao sistema elétrico e garantir a segurança dos equipamentos e dos operadores (??).

Contudo falhas de segurança durante as rotinas de um colaborador não são raras no setor elétrico, podendo causar danos à sua saúde, e em alguns casos levando a óbito. De acordo com o estudo de (??), acidentes de trabalho são muito comuns em ambientes como subestações de energia. Para avaliar as causas, foram analisados diversos processos trabalhistas durante um período de tempo contra empresas que prestam serviço neste setor. Foram concluídas as existência de várias as causas, mas uma se destaca: a falta de treinamento. No estudo citado, este fato é atrelado à terceirização dos serviços. Enquanto um funcionário direto da companhia de energia da região recebia 6 meses de treinamento, o funcionário terceirizado era treinado em um período médio de 30 a 40 dias. A qualidade do conteúdo destes treinamentos para o funcionário terceirizado também era muito mais superficial. Em uma análise dentro deste estudo, a respeito de um acidente fatal, foi verificado que um técnico foi acionado para resolução de um problema de rompimento de cabo. Devido a um descuido, um dos colaboradores tocou em um cabo energizado, sem se preocupar em verificar se todas chaves estariam desligadas, acabando por o levar a óbito. Deste estudo, portanto, entendeu-se que a busca por custos mais baixos durante o treinamento, pessoas que se expõe ao risco, e o próprio funcionamento da transmissão de energia é colocado à prova.

Nesse cenário, intervenções tecnológicas seriam de grande valia para melhorar a condição de treinamento de operadores no sistema elétrico. São diversas as possibilidades de aplicações que podem atuar nesse sentido, desde sistemas avançados de monitoramento

até soluções de Realidade Virtual (RV), oferecendo oportunidades para aprimorar a gestão e operação das subestações. A incorporação dessas inovações, podem levar as empresas do setor elétrico a oferecer maior segurança aos seus colaboradores, reduzir custos operacionais e garantir um fornecimento de energia mais confiável para os consumidores finais (??).

Para funções didáticas, como treinamentos, a RV se destaca como uma abordagem disruptiva em relação a métodos tradicionais, principalmente pelo alto nível de imersibilidade no contexto da aplicação, proporcionado ao usuário e ao alto resultado no aprendizado do conteúdo trabalhado. Contudo, para construir e preparar todo o ambiente para uma experiência imersiva em RV, faz-se necessário a elaboração de uma complexa estrutura que envolve desde a escolha do equipamento que será utilizado para a projeção ao usuário, como por exemplo, uma caverna de visualização, capacete de virtualização ou mesmo óculos de RV, até a criação, em softwares próprios para esse tipo de desenvolvimento, toda modelagem gráfica do ambiente até as interações que existentes na aplicação. Fatores como a capacidade gráfica e técnica são levadas em consideração nesta etapa, uma vez que aplicações com grande quantidade de interações e elementos, exigem do hardware que irá renderizar elevada capacidade de processamento. Se a demanda pela capacidade for alta, e não for suportada pelo hardware, será exigido do desenvolvedor redução na qualidade das texturas, assim como outros tratamentos para que toda a experiência durante a imersão não seja lenta ou mesmo careça de elementos que destituia a aplicação de imersibilidade (??).

Outro recurso que tem sido aplicado em várias áreas da ciência são as Redes Neurais Artificiais (RNA). Sua utilização tem sido atrelado a resolução de sistemas não-lineares em que nem todas as variáveis do problema são conhecidas, assim como problemas em que exista ruídos nos dados a serem tratados, ou seja, ideais para problemas do mundo real quando transportados para o mundo virtual. Ao simular o funcionamento do cérebro humano, replicando o aprendizado natural, as RNA exibem a capacidade de resolver problemas complexos, sendo, assim, ferramenta ótima a ser associada a um trabalho de pesquisa (??).

Deste modo, motivado pelo possibilidade de elaborar uma ferramenta que simplifique o desenvolvimento de um sistema em RV voltado para aplicações de treinamento de colaboradores em subestações de energia, este trabalho se propõe a construir uma ferramenta que faça a inserção automática de componentes em um ambiente de RV de uma subestação de energia. Toda a automação será alimentada por um modelo treinado a partir de uma RNA, alimentada por fotos capturadas por VANTs em duas subestações de energia diferentes.

1.2 Objetivos e Metas

O objetivo geral desta pesquisa é propor um sistema de inserção automática de componentes em ambientes virtuais de treinamento para subestações de energia a partir de imagens aéreas coletadas do local a ser mapeado virtualmente. Para alcançar esse objetivo geral, foram estabelecidos os seguintes objetivos específicos:

- Realizar uma revisão da literatura científica, para identificar quais os algoritmos utilizados no reconhecimento de padrões em imagens aéreas obtidas por VANTS;
- Estudar e avaliar quais são os hiperparâmetros do algoritmo de inteligência artificial a ser utilizado, que garanta maior eficiência no reconhecimento de componentes das subestações elétricas;
- Desenvolver uma automação capaz de receber uma imagem, reconhecer componente(s) da subestação elétrica inserir no ambiente virtual.

1.3 Estrutura da Dissertação

A presente dissertação é composta por sete capítulos, descritos da seguinte forma.

- No Capítulo 1 são apresentadas as motivações, os objetivos e a estruturação do trabalho;
- No Capítulo 2 são apresentados os principais fundamentos teóricos relacionados ao trabalho;
- No Capítulo 3 é apresentado o estado da arte da linha de pesquisa principal desse trabalho;
- Nos Capítulos 4 e 5 são apresentados materiais/métodos e detalhes de implementação;
- No Capítulo 6, são discutidos e apresentados os resultados obtidos nesse trabalho a partir do sistema desenvolvido;
- Por fim, no Capítulo 7, são apresentadas as conclusões e as perspectivas para trabalhos futuros.

2 Fundamentação Teórica

2.1 Subestações de Energia

Atualmente, o sistema elétrico do Brasil é formado por um conjunto de usinas, subestações, linhas de transmissão e demais equipamentos, que viabilizam a produção, transmissão e distribuição de energia elétrica. A energia é gerada em usinas que variam conforme os recursos utilizados, como hidroelétricas, termoelétricas, eólicas, nucleares, entre outras, e é então transportada por linhas de transmissão até subestações de energia de empresas distribuidoras. Nestas subestações, a tensão é ajustada para os níveis adequados de consumo. Esse processo elétrico é composto por três etapas principais: geração, transmissão e distribuição (Figura ??). O sistema elétrico é composto por diversos elementos, sendo os mais destacados os geradores, responsáveis pela transformação de energia mecânica em elétrica, e os sistemas de transmissão e distribuição, que conduzem a energia até os consumidores. As subestações têm um papel fundamental ao ajustar a tensão na transmissão e distribuição, assegurando a confiabilidade e a qualidade do serviço elétrico fornecido aos clientes (??).

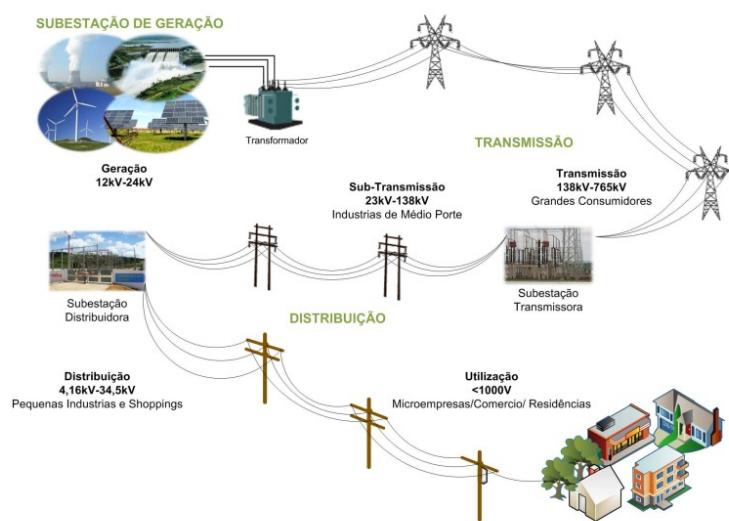


Figura 1 – Representação dos componentes do sistema de geração e transmissão de energia (??)

De modo geral, portanto, a energia é obtida por geradores e transmitida para subestações primárias. Há aí a primeira aferição de qualidade dos níveis de transmissão. Após isso, a energia transformada é enviada para subestações de distribuição, alterada para níveis mais baixos para serem transmitidas para setores industriais e urbanos (??).

É importante destacar que a energia produzida no Brasil é oriunda majoritariamente de fontes renováveis, em especial da geração hidroelétrica, correspondendo a 69% de toda malha produtora. Ela é considerada uma energia limpa, de baixo custo e possui reduzida emissão de gases de efeito estufa. (??).

Analizando mais detalhadamente as subestações, podemos afirmar que elas são encarregadas de regular a tensão tanto na transmissão quanto na distribuição de energia. Elas operam de forma autônoma, contando com sistemas supervisórios para o controle e supervisão remotos. Dentro das subestações, funções como comutação, proteção e controle também são desempenhadas. Um dos problemas que as subestações resolvem é a interrupção de curtos-circuitos, utilizando para isso disjuntores. Existem diversos tipos de subestações: as de transmissão, que conectam linhas de transmissão com diferentes voltagens; as de distribuição, que ligam linhas de transmissão a linhas de distribuição e regulam a tensão; e as subestações coletoras, que são utilizadas na geração de energia eólica para conectar a geração às linhas de transmissão. Portanto, as subestações são fundamentais para garantir a confiabilidade e a qualidade do fornecimento de energia (??).



Figura 2 – Reatores de Núcleo de ar em uma subestação de energia
(??)

Dentre os diversos equipamentos que compõem a subestação de energia, um em particular será objeto de estudo neste trabalho: os reatores de núcleo de ar (Figura ??). Contudo, o interesse reside em seu formato geométrico, que será tratado posteriormente. Nota-se que possui um corpo cilíndrico. Em seu topo, pode possuir ou não uma proteção circular contra intempéries. Estes dispositivos são essenciais para a regulação da reatividade e controle de correntes indutivas em sistemas de transmissão e distribuição de energia. Diferentemente dos reatores de núcleo de ferro, os reatores de núcleo de ar oferecem vantagens como uma resposta mais linear e a capacidade de operar sem saturação

magnética, o que os torna ideais para aplicações onde é necessário um comportamento previsível e estável (??).

2.2 Realidade Virtual

RV pode ser definida como um ambiente gerado a partir de um sistema computacional em que o usuário não apenas se sente dentro do contexto artificial, como também possibilita ao usuário a consciência de que pode navegar e interagir neste ambiente virtual. Trata-se, por isso, de uma interessante interface em um ser humano, um computador, capaz de gerar navegabilidade, interações e, principalmente, imersão em um ambiente sintético, gerado computacionalmente por meio de canais multisensoriais. Além disso, a RV pode ser classificada de duas formas diferentes: Imersiva e Não-imersiva. Na primeira, cria-se uma RV que isole o usuário do mundo real. Seus sentidos são bloqueados para recepção dos estímulos do seu entorno, para se direcionarem àqueles oriundos do mundo fictício. Para tanto, uma ampla variedade de equipamentos, como fones de ouvidos, luvas de dados, óculos de RV, entre outros, são empregados para os resultados esperados. Já a RV Não-Imersiva não se isola do mundo real. Ou seja, o usuário tem consciência de que está em um ambiente artificial. Da mesma forma, uma ampla variedade de equipamentos é empregada para gerar essa interação, incluindo dispositivos do cotidiano, como mouses, monitores de computador. Uma grande variedade de dispositivos convencionais e não-convencionais pode compor essa interação (??).

É importante que os três conceitos da RV sempre sejam presentes em uma aplicação do tipo, sendo eles: interação, imersão e navegação (??).

O surgimento da RV data de 1963, em que foi apresentado uma aplicação de manipulação de objetos tridimensionais em um computador, denominada Sketchpad (??). A aplicação conseguia reproduzir interatividade por meio de uma caneta óptica, que era utilizada para seleção de objetos projetados em uma tela. Neste trabalho, surgiram alguns dos principais termos da RV, como representação visual, dispositivos especiais, e interações em tempo real. Em 1968, o mesmo autor do trabalho anterior, Sutherland, publicou outro trabalho marcante para a história: A Head-Mounted Three Dimensional Display (??). Nele, foi introduzido o conceito de imersividade para uma RV; no caso, um capacete capaz de projetar fotos diretamente aos olhos do usuário, assim como rastrear o movimento da cabeça, afim de que este movimento fosse responsivo no ambiente virtual. Seguido a isso, uma série de outros equipamentos foram desenvolvidos a fim de sofisticar as soluções do ramo e propor novos modelos. A tendência sempre prevaleceu de surgir ferramentas mais simples e acessíveis ao usuário final (??).

Além das aplicações lúdicas, a RV atua de maneira séria e efetiva em diversas áreas técnicas. Como ferramenta para treinamento de operadores em ambientes de risco e de

difícil simulação, torna-se uma alternativa viável para capacitação. O trabalho de (??) demonstra bem essa ideia ao investigar técnicas de RV para que uma pessoa comum, sem treinamento anterior, possa, de maneira segura à uma subestação em funcionamento e à própria pessoa e todos ao seu redor, interagir com um transformador de energia, executando todos procedimentos que teria em um ambiente real, contudo de maneira virtual. Este treinamento já seria uma base interessante para que um profissional pudesse agilizar seu treinamento no mundo real ou mesmo absorvê-los. Em trabalhos de alta periculosidade tudo é crítico; então, qualquer intervenção que possa atenuar os riscos inerentes à natureza do trabalho, traz grandes avanços ao processo de profissionalização de técnicos e pessoas interessadas.

2.3 Redes Neurais Artificiais

A RNA trata-se de um conjunto de técnicas que buscam simular o funcionamento do cérebro humano, a partir de algoritmos computacionais, a fim de resolver problemas específicos. Sua eficiência está relacionada com a quantidade de interações que as unidades de processamento que o compõem realizam entre si. Comparam-se as RNA à mente humana devido a sua capacidade de aprendizado, podendo generalizar funções a partir de alguns exemplos informados, e delas prever o comportamento de valores para os quais não foi fornecido a resposta esperada. A base do funcionamento da RNA está no conceito do neurônio artificial (NA), que se traduz como uma pequena unidade de processamento, capaz de receber um sinal simples, e a partir dele gerar uma resposta. De acordo com (??), a primeira noção de NA surge em 1943, no trabalho Warren McCulloch e Walter Pitts, no artigo: “A Logical Calculus of the Ideas Immanent in Nervous Activity”.

Matematicamente, o NA recebe um valor de entrada, realizando com ele o produto desse valor a um outro denominado peso. O resultado é comparado com um diferenciador: se for maior, será dada a saída verdadeira; se menor, falso. Na Figura ??, esse processo é demonstrado de maneira esquemática. A operação descrita, chamada também de threshold logic, i.e., lógica limiar, em tradução livre, mimetiza o funcionamento de um componente eletrônico chamado transistor, base do funcionamento dos processadores computacionais modernos, em que a passagem de corrente elétrica por ele é interrompida ou permitida com base no sinal de entrada (??).

Aprofundando-se na álgebra relacionada do NA, pode-se visualizar o esquema do seu funcionamento na Figura ??, e também acompanhar a definição de cada envolvido nesta operação (??):

- **Sinais de entrada:** Constituem-se os sinais, ou valores, externos ao modelo, muitas vezes submetidos a algum tratamento prévio, responsáveis por alimentar a rede.

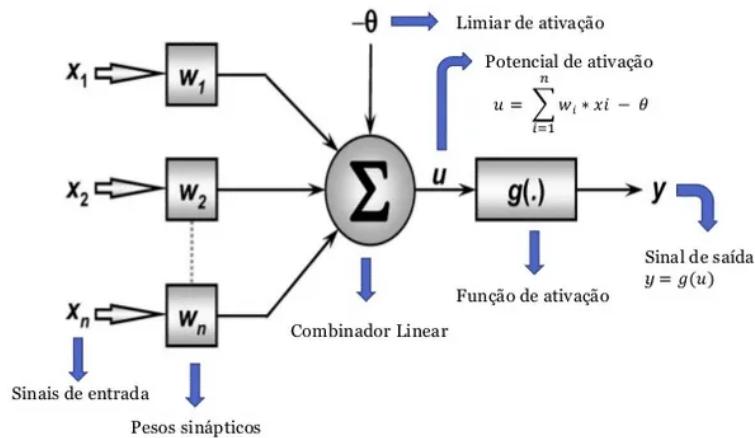


Figura 3 – Representação de um NA (??)

- **Pesos sinápticos:** Também chamados apenas de pesos, ponderam os sinais de cada entrada da rede.
- **Combinador linear:** Operação aritmética envolvendo os sinais de entrada a fim de gerar um potencial de ativação.
- **Bias:** O bias é um valor adicional que é somado à combinação linear das entradas antes de passar pela função de ativação, visando ajustá-la para que os dados melhor se adaptem à rede.
- **Limiar de ativação:** Também denominado threshold, determina o nível adequado em que o resultado obtido pelo combinador linear pode acionar a ativação.
- **Potencial de ativação:** É o resultado decorrente da discrepância entre o valor gerado pelo combinador linear e o limiar de ativação. Se esse resultado for positivo, indicando $u \geq 0$, o neurônio gera um potencial de excitação; caso contrário, o potencial será inibitório.
- **Função de ativação:** Sua função é restringir a saída de um neurônio dentro de um determinado intervalo de valores.
- **Sinal de saída:** Resultado final da saída, que pode ser utilizado como entrada para outros neurônios interligados sequencialmente.

Em suma, a RNA, enquanto um conjunto de NA, pode ser entendido como um processador robusto, distribuído de maneira paralela, com pequenas unidades de processamento. Sua semelhança ao cérebro humano, deve-se, portanto, à capacidade de aprendizado e aos sinais sinápticos, existentes entre os neurônios, responsáveis pelo processo de armazenamento de conhecimento (??).

De acordo com (??), em 1958, após à concepção do NA, a primeira RNA a se notabilizar e a continuar sendo utilizada até o presente momento trata-se da Perceptron (??). Este algoritmo possuía a capacidade de alterar os pesos dos neurônios, conforme o avanço do processamento da rede, de modo a resolver problemas da classificação linear. Seguido a ela, em 1960, houve a concepção da rede Adaline (??). Diferentemente da perceptron, esta rede já se mostrava capaz de produzir resultados que iam além dos valores binários, sendo capaz de gerar respostas de valores presentes em todo o conjunto dos números reais. O cálculo da rede Adaline era baseado na regra Delta, que era capaz de aproximar os valores dos pesos gerados para aqueles valores com menor erro possível. Ambas as redes, contudo, apresentavam uma limitação quanto à modificação dos pesos em todas suas multicamadas. Para esse cenário, foi proposto o conceito de backpropagation, ou, em tradução livre, retropropagação do erro (??). A ideia da retropropagação se inicia com o feedforward, em que todos os dados de entrada de uma RNA são transmitidos do início à camada de saída, sem nenhuma alteração de peso. Desta forma, calcula-se o erro total dessa primeira passagem, comparando o valor resultante com o esperado. Essa métrica de erro torna-se o guia para o ajuste dos pesos entre as conexões de cada neurônio. Com ela, é feito o caminho de volta, e calculado um gradiente de erro, que será o fator decisivo para a atualização dos pesos durante todo o treinamento.

A topologia básica de uma rede neural é dividida em três níveis de camada: a camada de entrada, a camada oculta e a camada de saída, conforme mostrado na Figura ???. A camada de entrada é, naturalmente, aquela que recebe os dados de entrada. Nela, cada nó representa uma característica ou atributo dos dados que estão sendo alimentados na rede neural. Por exemplo, em uma aplicação de reconhecimento de imagens, cada nó na camada de entrada pode representar um pixel da imagem. A camada oculta, que não se restringe a apenas uma, podendo ser várias, representa um processo intermediário entre a camada de entrada e a camada de saída. Cada neurônio em uma camada oculta recebe entradas das camadas anteriores, realiza algum tipo de transformação não linear dessas entradas e passa o resultado para a próxima camada. A presença de múltiplas camadas ocultas permite que a rede aprenda representações complexas e abstratas dos dados. Por fim, a camada de saída representa a camada final, e ela é responsável por gerar as saídas desejadas. A estrutura da camada de saída depende do tipo de problema que está sendo resolvido. Por exemplo, em um problema de classificação, cada nó na camada de saída pode representar uma classe diferente, e a saída pode ser interpretada como a probabilidade de pertencer a cada classe (??).

Essencialmente, as redes neurais aprendem iterativamente ajustando os pesos de suas conexões através do processo de treinamento, onde são apresentados a um conjunto de dados de entrada e as correspondentes saídas desejadas. Com o tempo, a rede neural é capaz de aprender a mapear efetivamente os padrões nos dados de entrada para as saídas desejadas, tornando-se assim capaz de realizar tarefas como reconhecimento de padrões,

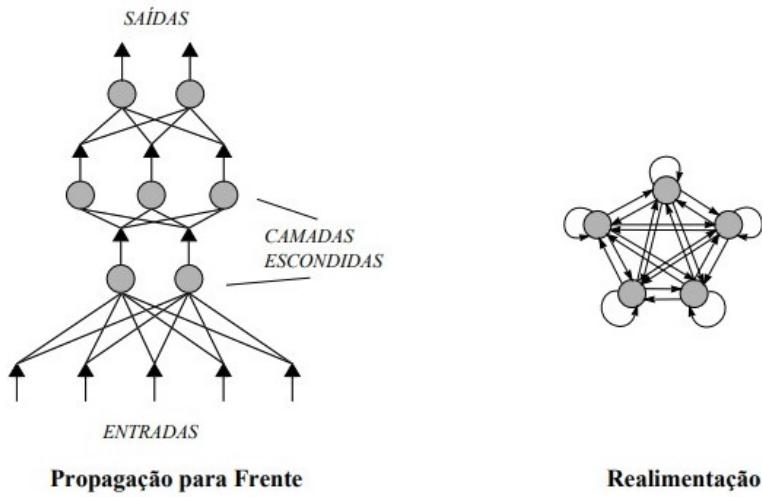


Figura 4 – Topologia básica de uma rede neural (??)

classificação, regressão, entre outros (??).

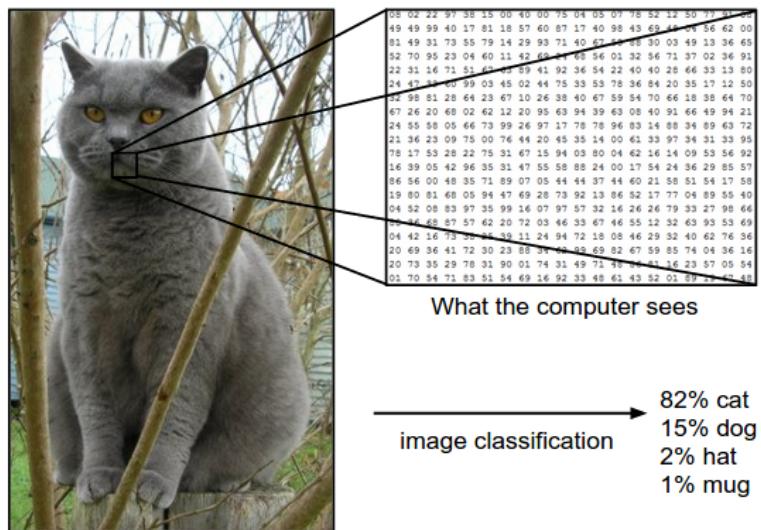


Figura 5 – Exemplificação de como uma imagem nada mais é que uma matriz, e como a filtragem busca padrões pré-estabelecidos dentro da imagem (??)

Contudo, para problemas envolvendo aprendizado por meio de imagens, a forma de lidar com as informações é diferente. De fato, imagens são dados, organizados em formato de matrizes de duas dimensões ou três dimensões (se for considerada a camada de cores), em que cada unidade de informação se chama pixel, conforme se nota no exemplo da Figura ???. Para processar esses dados, um tipo de RNA destaca-se: a Rede Neural Convolucional (RNC). Comparando ambas, nota-se que a RNA é composta por camadas intrinsecamente conectadas, em que cada neurônio de uma camada se conecta a todos os neurônios da camada seguinte. Isso faz com que ela seja adequada para dados vetorizados e para problemas de classificação e regressão sem uma estrutura espacial ou temporal específica. Por sua vez, a RNC recorre a dois tipos de camadas (Figura ??): a convolucional,

cujo objetivo é extrair características locais, e a de pooling, responsável por manter a estrutura espacial essencial (??).

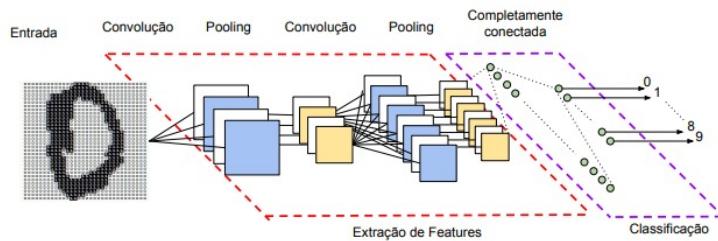


Figura 6 – Esquema de funcionamento de uma RNC (??)

A camada de convolução, portanto, é responsável pela feature extractor, i.e., pela extração de características de interesse em uma imagem. Basicamente, são aplicados filtros, também chamados de kernel, à imagem a fim de buscar padrões. São comumente empregados filtros de detecção de bordas (edge detection), desfoque (blur), nitidez (sharpen). Conforme na Figura ??, cada elemento do filtro é multiplicado pelo elemento de mesma posição na região em que o filtro está sendo aplicado naquele instante. Ao final dessas operações matriciais, adiciona-se os resultados dessas multiplicações para ter um único valor como saída, que será o pixel correspondente na imagem filtrada (??) .

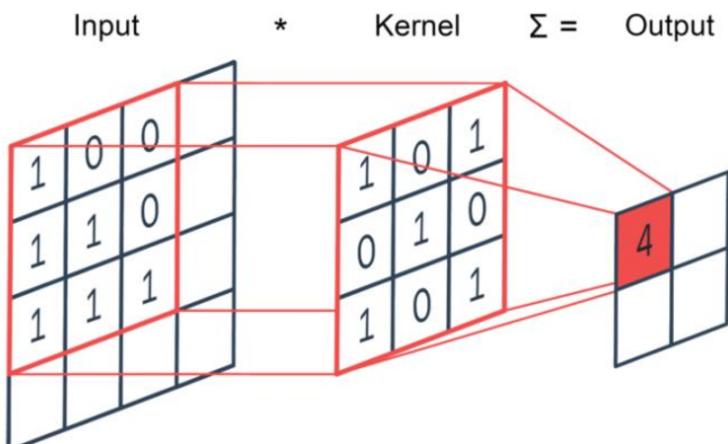


Figura 7 – Operação de filtragem na camada convolucional. Nota-se que o filtro é uma multiplicação de matrizes (??)

A camada de pooling é aplicada após cada camada convolucional, com o objetivo de reduzir as dimensões das imagens enquanto mantém a profundidade do volume. Esse processo promove a invariância da Rede Neural Convolutinal (RNC) às transformações geométricas, permitindo à rede reconhecer objetos na imagem independentemente de sua posição. Além disso, o uso da camada de pooling contribui para diminuir significativamente o custo computacional da rede. Existem diferentes tipos de pooling, como o max pooling, que seleciona o valor máximo em uma região específica da imagem, e o average pooling,

que calcula a média dos valores nesta região. O max pooling é o mais comum, pois preserva as características mais salientes das imagens, enquanto reduz o ruído. A camada de pooling, ao reduzir o tamanho espacial dos mapas de características, também ajuda a evitar um problema comum no aprendizado de máquina, em que um modelo aprende não apenas os padrões relevantes dos dados de treinamento, mas também o ruído e as particularidades específicas desses dados. Chama-se este problema de overfitting. Evitando-o, a rede consegue construir um treinamento mais generalizável e que possa gerar resultados relevantes para dados diferentes do modelo treinado (??).

Por último, a camada totalmente conectada emprega uma função de ativação na sua camada de saída para realizar a classificação. Ser "totalmente conectada" significa que todos os neurônios da camada anterior estão ligados a todos os neurônios da camada subsequente. Ela recebe as saídas da camada de convolução e da camada de pooling e, em seguida, utiliza essas informações para determinar a classe à qual a imagem de entrada pertence (??).

As RNC representam por si só uma revolução no campo da visão computacional, tornando-se a base para uma variedade de aplicações, desde reconhecimento de imagens até análise de vídeos. Entre as arquiteturas mais notáveis estão a LeNet, AlexNet, VGGNet, GoogLeNet e ResNet, cada uma contribuindo com avanços significativos em profundidade, eficiência e precisão. Além dessas, uma destaca-se em especial: a YOLO ("You Only Look Once"), uma abordagem inovadora para treinamento e detecção de objetos (??).

2.3.1 You Look Only Once

YOLO, uma Rede Neural Convolutiva (RNC), é conhecida por sua eficácia e precisão na detecção de objetos em imagens e vídeos. Desenvolvida por Joseph Redmon, Santosh Divvala, Ross Girshick e Ali Farhadi em 2015, a arquitetura do YOLO foi apresentada em sua primeira versão no artigo "You Only Look Once: Unified, Real-Time Object Detection"(??).

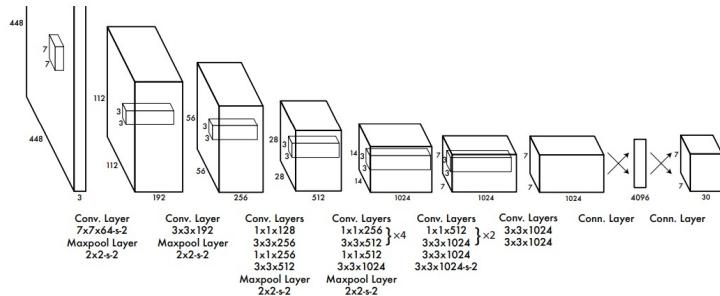


Figura 8 – Disposição das 24 camadas convolucionais e das 2 camadas totalmente conectadas (??)

A arquitetura do YOLO é singular: ao contrário de outras RNCs que fazem múltiplas

passagens pela imagem em busca de objetos, o YOLO analisa a imagem inteira de uma vez, justificando seu nome, que em tradução livre significa “você olha apenas uma vez”. Essa abordagem permite ao YOLO aplicar uma única rede neural à imagem completa. Durante o processamento, a imagem é dividida em regiões menores, e a rede prevê caixas delimitadoras, as probabilidades de existência de objetos nessas caixas, e a classe provável de cada objeto. Sua classificação, enquanto rede neural, vai além da RNC, pois ela utiliza uma arquitetura conhecida como Darknet, um tipo de rede neural profunda (variação da RNC) e sua implementação original foi desenvolvida em C, embora agora esteja disponível em várias outras linguagens de programação, graças ao apoio da comunidade e de empresas. Quando utilizada para treinamento sua estrutura básica, estabelecida em sua primeira versão, segue, evidentemente, o padrão de uma RNC. As camadas convolucionais iniciais da rede extraem características da imagem, enquanto as camadas totalmente conectadas preveem as probabilidades de saída e as coordenadas (Figura ??). Em sua primeira versão, a rede possuia 24 camadas convolucionais seguidas por 2 camadas totalmente conectadas. citeyoloVisaoComputacional. Embora existam várias versões do YOLO, há um funcionamento geral da arquitetura. O primeiro passo do YOLO é dividir a imagem em uma grade de células S por S . Nas versões iniciais, essa grade era de 13×13 , totalizando 169 células, enquanto nas versões mais recentes é de 19×19 . Cada célula é responsável por prever/detectar 5 caixas delimitadoras, pois pode haver múltiplos objetos em uma única célula. Portanto, na versão de exemplo do YOLO, há um total de 845 caixas delimitadoras ($13 \times 13 \times 5$). Na Figura ??, esse processo é representado. Notam-se diversas caixas, aquelas com bordas mais grossas indicam que a probabilidade de a demarcação estar representando o objeto que se propõe é alta. (??).

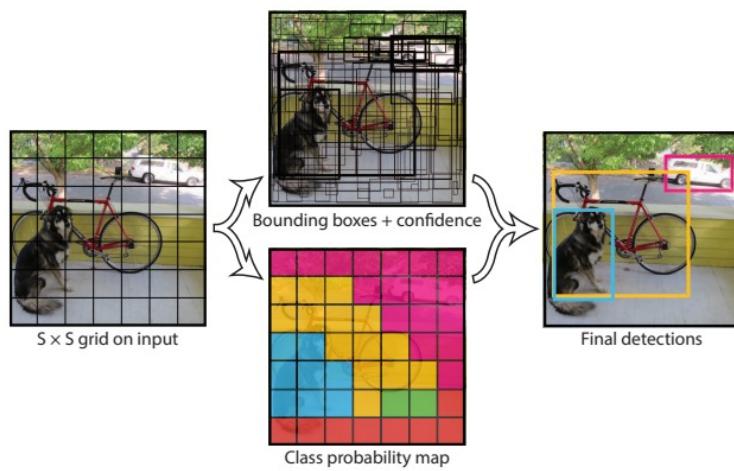


Figura 9 – Processo de predição da Yolo (??)

A partir da YOLOv3, foram introduzidos conceitos para definir algumas partes de sua arquitetura. Primeiramente, define-se a camada de backbone, ou espinha dorsal. Basicamente, é nesta etapa que ocorre a extração de características significativas para o

processo de classificação. Ela captura características hierárquicas em diferentes escalas, com características de nível inferior (como bordas e texturas) nas camadas iniciais e características de nível superior (como partes de objetos e informações semânticas) nas camadas mais profundas. O neck, ou pescoço, conecta a camada backbone à head, ou cabeça, agregando e refinando as características extraídas, muitas vezes focando em melhorar a informação espacial e semântica em diferentes escalas. A head processa as características fornecidas pelo neck, gerando previsões para cada candidato a objeto. Resumindo, a partir da YOLOv3, a arquitetura foi dividida em três partes: backbone para extração de características; neck para agregar e refinar características; e head para fazer previsões. A espinha dorsal usa uma CNN para capturar características hierárquicas, o pescoço melhora a informação espacial e semântica, e a head gera as previsões finais, que são refinadas por pós-processamento para eliminar sobreposições (??). Atualmente, a oitava versão da YOLO, ou, YOLOv8 , é considerada estado da arte em se tratando de análise e treinamento de imagens (??). YOLOv8 traz uma série de avanços significativos em relação às versões anteriores, destacando-se pela melhoria no desempenho e precisão das detecções. A arquitetura foi otimizada para equilibrar velocidade e precisão, utilizando backbones modernos como CSPDarknet para uma extração de características mais eficiente. O processo de treinamento foi acelerado e aprimorado com melhores métodos de regularização e otimização, enquanto a facilidade de uso e integração foi aprimorada, atendendo às demandas tanto de dispositivos de alta performance quanto de dispositivos móveis com menor capacidade de processamento (??).

2.3.2 Batch

Um dos aspectos cruciais do funcionamento da YOLO é o conceito de "batch" (em tradução livre, "lote") durante o treinamento da rede neural. Ao agrupar várias imagens em lotes para processamento simultâneo, a YOLO aproveita a capacidade de processamento paralelo das GPUs, acelerando significativamente o treinamento. Durante a propagação direta, cada imagem no lote é processada pela rede neural para gerar previsões de detecção de objetos. Em seguida, a perda é calculada em relação às anotações verdadeiras, e os pesos da rede são atualizados para minimizar essa perda, usando algoritmos de otimização como o gradiente descendente. Esse processo é repetido para vários lotes de imagens até que a rede converja para uma solução adequada. Assim, o uso eficiente de lotes na YOLO não apenas acelera o treinamento, mas também contribui para a robustez e eficácia dos modelos de detecção de objetos resultantes. (??)

2.3.3 Otimizadores

O objetivo dos algoritmos de aprendizado de máquina supervisionados é otimizar a redução de uma função objetivo, geralmente chamada de função de custo J. Reduzir

essa função é crucial, pois a rede precisa aprender os pesos que melhor representam o relacionamento entre os dados, formando um modelo preditivo para fazer previsões em novos conjuntos de dados. O algoritmo de otimização mais simples para encontrar esses pesos é o Gradiente Descendente (GD), onde se realiza pequenos passos em direção ao mínimo global da função de custo. No entanto, o GD enfrenta dificuldades com conjuntos de dados muito grandes devido à sua abordagem em lote, tornando o treinamento computacionalmente caro. Uma alternativa popular é o Gradiente Descendente Estocástico (SGD, sigla para a expressão em inglês “Stochastic Gradient Descent”), que aplica atualizações de peso com base em amostras aleatórias de dados de treinamento, resultando em convergência mais rápida. O SGD possui três parâmetros importantes que podem influenciar no treinamento da rede neural: a taxa de aprendizagem, que determina quão rápido o otimizador tentará treinar a rede neural, com uma taxa muito alta podendo levar a falhas no treinamento e uma taxa muito baixa resultando em treinamento lento; o momentum, que indica quanto a direção da mudança de peso anterior deve influenciar na etapa atual, acelerando o treinamento ao permitir que o otimizador "persista" em direções com histórico de melhoria; e o decay, utilizado para diminuir a taxa de aprendizagem conforme os erros diminuem durante o treinamento, evitando oscilações excessivas e permitindo uma convergência mais suave do modelo.(??).

Além do SGD, outro algoritmo de otimização recorrente é o Adam (“Adaptive Moment Estimation”, em tradução livre, “Estimação Adaptativa de Momento”). O otimizador Adam é um algoritmo popular de otimização de gradientes utilizado em treinamento de redes neurais e outras tarefas de aprendizado de máquina. Ele calcula e mantém estimativas adaptativas dos momentos do gradiente, incluindo o primeiro momento (média) e o segundo momento (variância). Essas estimativas são utilizadas para calcular uma correção de gradiente, chamada de momento Adam, que ajusta adaptativamente a taxa de aprendizagem para cada parâmetro. Além disso, o Adam inclui termos de regularização L2 para evitar que os pesos cresçam excessivamente durante o treinamento. Essa abordagem combinada de momentos adaptativos e ajuste de taxa de aprendizagem torna o Adam eficiente, fácil de implementar e eficaz em uma variedade de problemas de otimização (??). Há também uma variação do Adam, que seria o AdamW. Esta variante do otimizador Adam que inclui regularização de peso ponderada diretamente no processo de atualização dos parâmetros, ao contrário do Adam tradicional, onde a regularização é aplicada separadamente após a atualização. Isso ajuda a corrigir problemas de "decaimento de pesos" em modelos de aprendizado profundo, resultando em melhor desempenho e capacidade de generalização (??).

2.3.4 Precisão e Recall

A fim de avaliar o desempenho de um treinamento na arquitetura YOLO, é preciso entender os resultados fornecidos pelo modelo. De acordo com (??), basicamente a YOLO utiliza a métrica chamada Average Precision (AP) (“Precisão Média”, em tradução livre). Ela se baseia no conceito de IoU (“Intersection over the Union”, Intersecção sobre a União, em tradução livre), que calcula uma razão entre a interseção da detecção feita pelo algoritmo com relação à marcação (bounding box) realizada em cima da área dividida pela união dessas duas áreas (Figura ??). Essa razão poderá ser comparada com um valor pré-estabelecido, o thresholds, que será referido por L. A partir desse valor, é possível que no processo de detecção retorne três diferentes resultados na pesquisa pela classe desejada. São eles: Verdadeiro Positivo (VP), Falso Positivo (FP) e Falso Negativo (FN). VP trata-se dos resultados considerados corretos que a rede neural retorna, que seriam todos resultados que $\text{IoU} > L$. FP já seriam os resultados em que $\text{IoU} < L$, que são tidos como incorretos. FN, por sua vez, trata dos resultados totalmente fora do esperado.

$$\text{Recall} = \frac{VP}{VP + FN} \quad (1)$$

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (2)$$

Com esses valores, pode-se calcular os resultados da saída que o treinamento da rede YOLO fornece. Em (1), tem-se o cálculo da Recall, que calcula a capacidade da rede de detectar todos os objetos relevantes em uma imagem. Já a Precisão (2), refere-se à capacidade da rede de encontrar apenas resultados relevantes.

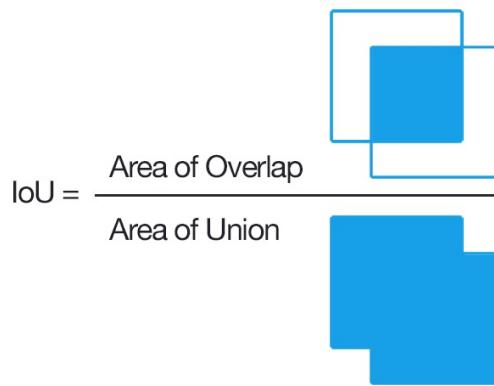


Figura 10 – Cálculo de IoU. (??)

A arquitetura YOLO disponibiliza um dataset, ou seja, um banco de imagens e weights, comum em todas as versões, chamado de COCO (“Common Objects in Context”, que em tradução livre seria “Classes Comuns de Objetos”) com classes pré-treinadas e imagens para realização de treinamentos. A partir dele, verificou-se por meio de testes a

eficiência das quatro últimas versões da YOLO, a fim de identificar se nas mais recentes houve melhorias significativas em termos de performance e precisão. Na Figura ??, é apresentado o comparativo das versões. Nota-se que a v8, para um menor número de parâmetros que as demais, apresentou uma mAP50-95, maior que nas versões v5, v6 e v7. Além disso, com relação a velocidade de processamento, a v8 também se sobressai, com maior rapidez no processamento, ao processar maior quantidade de imagens para um mesmo intervalo de tempo (??).

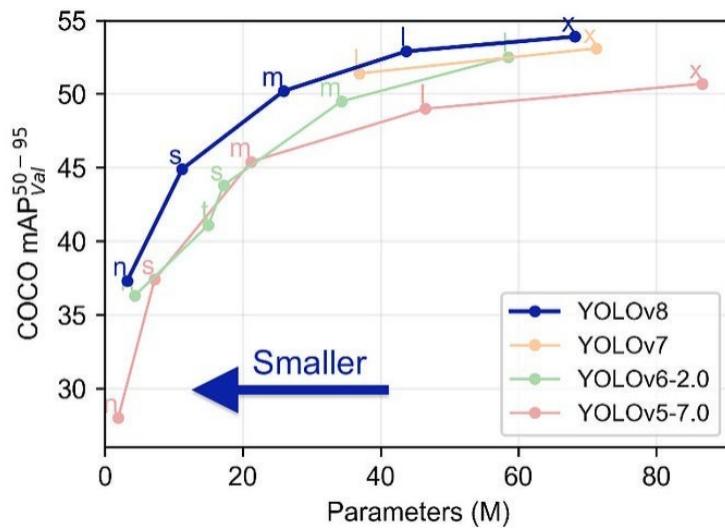


Figura 11 – Aumento de desempenho da precisão média, versão 5 para 8. (??)

2.4 Considerações finais

Neste capítulo, foi apresentado o arcabouço teórico necessário para o entendimento da proposta dessa dissertação. A apresentação das redes neurais, e em específico, o modo que a arquitetura da YOLOv8 sobrepõe-se em termos de eficiência em relação às demais arquiteturas abordadas em outros trabalhos científicos, demonstra o direcionamento assertivo deste trabalho. Além disso, a apresentação da Realidade Virtual como uma disciplina inovadora e muito útil para servir à diversos propósitos dentro da indústria e ciência, corroboram para o entendimento da proposta desta dissertação.

3 Trabalhos Relacionados

3.1 Introdução

Este capítulo apresenta os trabalhos relacionados à utilização de RNC para treinamento de imagens, com e sem variação de parâmetros; a trabalhos que aplicam especificamente a versão 8 do YOLO; aqueles que utilizaram VANT e os que criaram automação para sistemas de Realidade Virtual. O objetivo é encontrar o estado da arte de cada trabalho, destacando as principais contribuições.

No final do capítulo, os estudos são comparados para justificar o sistema abordado neste trabalho.

3.2 Alteração de Parâmetros da RNC

3.2.1 Identificação e Medição de Defeitos em Produtos Automotivos Usando Visão Computacional

A dissertação apresentada por (??) investigou a aplicabilidade e eficiência de um sistema de visão computacional para identificar defeitos visuais no controle de qualidade de peças automotivas reposição. O intuito do trabalho foi de avaliar a possibilidade de automatizar de modo confiável o processo de inspeção e precificação do reparo destas peças.

Para este trabalho, foi utilizado um dataset composto por imagens de vidros automotivos, fornecidos por uma empresa do ramo. Os defeitos analisados incluíram bolhas, delaminação, irisação, ostra e grau em vidros, além de manchas em peças como faróis, lanternas e retrovisores. Todos esses defeitos foram identificados nas fotos fornecidas. Assim como na presente dissertação, foi utilizado o software LabelImg para realizar a marcação da ocorrência de cada uma dessas anomalias no conjunto de fotos, como se vê na Figura ??.

Em sua fundamentação, foram entendidas as RNC como ferramentas eficazes para o processamento de imagens. Avaliou-se a rede ResNet (??), como uma possibilidade de arquitetura para processar as imagens, justamente por se mostrar uma ferramenta poderosa para a identificação de padrões. Para tanto, ela se vale de uma série de camadas residuais com conexões de salto, permitindo que os gradientes fluam mais facilmente através da rede durante o treinamento, mitigando o problema do desaparecimento do gradiente. Esse problema ocorre quando, em redes profundas, os gradientes se tornam



Figura 12 – Exemplo de marcação de foto utilizando labelImg, de um defeito em um vidro (??)

extremamente pequenos ao retropropagar através das camadas, dificultando a atualização dos pesos nas camadas iniciais e, consequentemente, a aprendizagem adequada da rede. A ResNet supera essa dificuldade utilizando conexões que pulam uma ou mais camadas, somando a entrada diretamente à saída dessas camadas puladas, permitindo a construção de redes muito mais profundas sem a degradação do desempenho (Figura ??). Com isso, a ResNet consegue capturar e aprender padrões complexos presentes nas imagens, resultando em uma melhoria significativa na precisão das tarefas de classificação e reconhecimento de imagens.

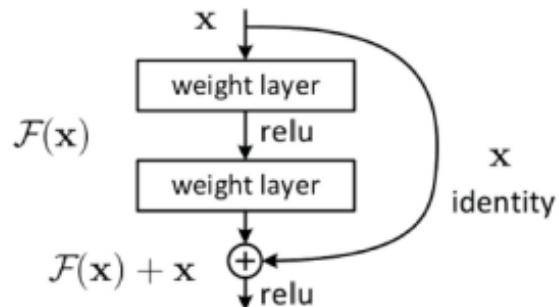


Figura 13 – Bloco de aprendizado da ResNet (??)

Contudo, a ResNet, devido à sua grande quantidade de camadas e processamento, acaba por ser morosa durante o treinamento. Além disso, sua utilização não é trivial e exige configurações específicas. Deste modo, o trabalho explora as vantagens da YOLO como alternativa. A YOLO é extremamente rápida, treinando em imagens completas com uma única passagem pelas imagens, o que reduz significativamente o tempo de processamento. Sua arquitetura simples e eficiente permite a detecção de múltiplos objetos em uma única passagem pela rede, tornando o treinamento mais direto e eficaz em termos de recursos computacionais. Além disso, a YOLO é altamente generalizável e menos propensa a erros

de background, sendo uma solução mais prática e veloz para detecção de objetos.

Para o treinamento, foram analisadas 3.397 imagens com defeitos específicos da rotina da empresa. Destas, 70% foram utilizadas para treinamento e 30% para validação. Todas as imagens foram anotadas manualmente. Em cada uma delas, pelo menos um dos defeitos estudados deveria constar nela para que a imagem fosse colocada para treinamento. Essa parte do processo garantiu que o modelo gerado fosse significativo.

Além disso, foram variados os tamanhos do valor do batch size (entre 2 e 32) e testados três otimizadores diferentes: SGD, Adam e AdamW. A YOLOv5 oferece variações da arquitetura para diferentes propósitos. Neste trabalho, foram testadas 10 variações da arquitetura, juntamente com a alteração de parâmetros. Todos os outros hiperparâmetros foram mantidos na configuração padrão. Cada treinamento foi realizado em 300 épocas.

Na figura (Figura ??), é apresentado o resultado do treinamento, com os respectivos valores, com a variação de parâmetro. Nota-se, disso, que o otimizador SGD, na variação YOLOv5x, com batch size de 8, alcançou a melhor precisão média, com mAP de 0,72921, mostrando com isso a melhor configuração. Com esse treinamento aplicado na identificação de imagens, resultou-se em 83,33% de precisão na especificação correta dos produtos defeituosos, mostrando com isso que a automação foi bem sucedida em seu propósito.

	OTIMIZADOR														
	Adam				AdamW				SGD						
A YOLOv5n	0,27538	0,34865	0,38961	0,41080	0,42883	0,43552	0,49647	0,57268	0,55400	0,56654	0,64075	0,65211	0,64726	0,66364	0,66875
R YOLOv5s	0,34196	0,36912	0,39459	0,44227	0,48022	0,48880	0,57665	0,59021	0,60678	0,60352	0,65081	0,68531	0,67008	0,68049	0,71152
Q YOLOv5m	0,29176	0,29275	0,39649	0,39963	0,44942	0,46095	0,58621	0,60273	0,61812	0,61064	0,69882	0,69744	0,71047	0,70489	0,68441
U YOLOv5l	0,22021	0,27639	0,35265	0,39097	-	0,40879	0,40433	0,52815	0,55821	-	0,68594	0,70907	0,71805	0,70637	-
I YOLOv5x	0,23398	0,29198	0,33194	0,35104	-	0,51829	0,53109	0,40974	0,60228	-	0,71413	0,71037	0,72921	0,70505	-
T YOLOv5n6	0,26231	0,31297	0,41337	0,44406	0,47405	0,44321	0,48123	0,57403	0,61179	0,59240	0,71131	0,70117	0,69670	0,70092	0,70257
E YOLOv5n6	0,22994	0,30712	0,34329	0,44475	-	0,46749	0,50013	0,58997	0,62237	-	0,71564	0,70941	0,71138	0,72084	-
T YOLOv5s6	0,22994	0,30712	0,34329	0,44475	-	0,46749	0,50013	0,58997	0,62237	-	0,71564	0,70941	0,71138	0,72084	-
U YOLOv5m6	0,28451	0,34191	0,32272	-	-	0,46582	0,49921	0,51498	-	-	0,70402	0,71544	0,72612	-	-
R YOLOv5l6	0,30023	0,39947	-	-	-	0,49962	0,56281	-	-	-	0,71112	0,71016	-	-	-
A YOLOv5x6	0,32116	0,38698	-	-	-	0,51172	0,55710	-	-	-	0,70572	0,71120	-	-	-
	Batch 2	Batch 4	Batch 8	Batch 16	Batch 32	Batch 2	Batch 4	Batch 8	Batch 16	Batch 32	Batch 2	Batch 4	Batch 8	Batch 16	Batch 32

Figura 14 – Resultado do treinamento variando parâmetros

3.3 Aplicação YOLOv8

3.3.1 Comparação de Modelos YOLOv5 e YOLOv8 para Detecção de Objetos em Áreas Rurais Usando Transferência de Aprendizado

O artigo de (??), apresenta um estudo comparativo entre duas versões da YOLO, a YOLOv5 e YOLOv8. Nele, avalia-se o desempenho de ambos os modelos na detecção de objetos em cenários característicos de ambientes rurais, onde os desafios incluem a presença de objetos agrícolas, plantações, animais e estruturas. Além disso, é utilizada uma técnica de transferência de aprendizado com modelos pré-treinados para adaptar ambas arquiteturas ao contexto proposto.

Como embasamento teórico, buscou-se um estudo similar ao apresentado, em que foi realizada também uma comparação dos modelos YOLOv5 e YOLOv8 para detecção

de veículos e placas em sistemas de transporte inteligentes, usando transferência de aprendizado com dados da plataforma Kaggle (??). Após uma avaliação abrangente, concluíram que o YOLOv8 superou ligeiramente o YOLOv5, além de ter um tempo de treinamento menor.

Já no trabalho citado de (??), foi estudada a detecção e contagem de plantas usando inteligência artificial, aplicando modelos de visão computacional para detectar e contar eucaliptos em plantações. O modelo R-CNN Resnet 101 alcançou 95% de precisão com um tempo de inferência de 578 milissegundos por imagem, destacando-se como o mais promissor para o desenvolvimento de software automatizado na silvicultura.

Por fim, citando (??), foi explorada a detecção de objetos com a arquitetura YOLO, comparando várias versões do modelo usando o conjunto de dados COCO. O YOLOv5 se destacou, alcançando uma acurácia de 67,9% na métrica mean-average Precision (mAP), superando significativamente as versões anteriores.

A literatura indica que a detecção de objetos tem atraído muita atenção, com muitos estudos focando na identificação e avaliação dos melhores modelos conforme o cenário de aplicação. Este trabalho é similar ao estudo (??) ao comparar YOLOv5 e YOLOv8, mas difere na área de aplicação.

Em sua conceituação, reforça-se a definição de objeto, a entidade desejável de ser reconhecida pela RNC. É ele qualquer forma visual reconhecível em uma imagem, como carros, pessoas, animais, ou prédios. Cada objeto será representado por uma classe por sua vez é representado por uma classe. Assim, a principal tarefa da detecção de objetos de uma RNC é identificar e localizar a presença dessas classes em imagens ou vídeos.

A metodologia deste trabalho envolveu a coleta de 452 imagens de 1800 pixels, utilizando um VANT. As classes definidas foram: cafezal, milharal, soja, estrada, casa, carro e pasto. As imagens foram anotadas usando a plataforma Roboflow para criar caixas delimitadoras. O pré-processamento incluiu orientação automática e redimensionamento das imagens para 640 x 640 pixels. O aumento de dados foi realizado com rotações aleatórias e adição de ruído, ampliando o conjunto para 1100 imagens. O dataset foi dividido em treinamento (960 imagens), validação (95 imagens) e teste (45 imagens), e exportado nos formatos "YOLO v5 PyTorch" e "YOLO v8".

No estudo, foi utilizado para complementar, pesos pré-treinados no conjunto MS COCO, com ajuste fino dos modelos para o novo conjunto de dados. A taxa de aprendizado foi de 0,0001 para YOLOv8x e 0,00001 para YOLOv5x, com um batch size igual a 128. Assim como nos demais trabalhos, foram adotadas as métricas de precisão, recall e mAP.

Os resultados deste estudo demonstraram que o modelo YOLOv8x superou o YOLOv5x em termos de precisão e eficiência na detecção de objetos em áreas rurais. Especificamente, o YOLOv8x alcançou maior acurácia nas predições e menor tempo total

de inferência, apesar de necessitar de um tempo de treinamento ligeiramente superior.

O YOLOv8x mostrou-se mais eficaz na detecção de objetos, atingindo um mAP de 0,767 (Figura ??). Além disso, apresentou maior confiança nas detecções, com um tempo médio de inferência de 15,9 ms. Já a YOLOv5x, por sua vez, atingiu um mAP de 0,735 (Figura ??), com um desempenho inferior em comparação ao YOLOv8x tanto em precisão quanto no tempo total de inferência. O tempo médio de inferência para o YOLOv5x foi de 17,2 ms.

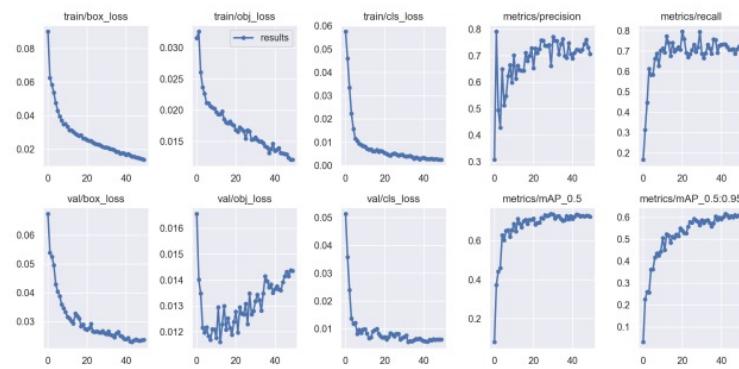


Figura 15 – YOLOv5 (??)

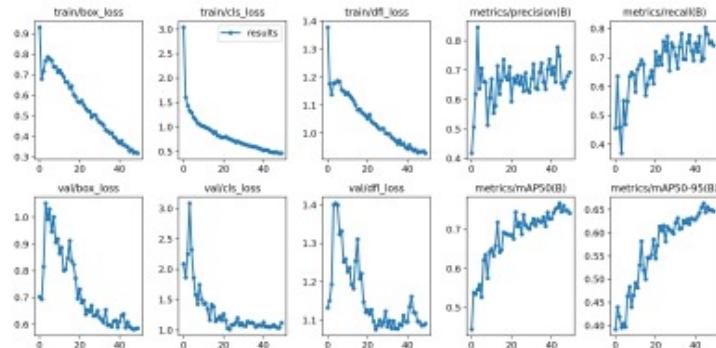


Figura 16 – YOLOv8 (??)

Este trabalho investigou a importância da detecção de objetos em imagens de áreas rurais, comparando os modelos YOLOv5 e YOLOv8. A detecção precisa de objetos é crucial em várias aplicações agrícolas, de monitoramento e preservação ambiental, aumentando a eficiência, produtividade e sustentabilidade nessas áreas. Os modelos YOLOv5 e YOLOv8 destacaram-se na detecção de objetos em tempo real, com o YOLOv5 sendo eficiente no treinamento e o YOLOv8 se sobressaindo pela precisão na detecção.

3.3.2 UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios

O artigo de (??) se inicia apresentando o crescente uso de VANTs para diversas aplicações da ciência e engenharia. Contudo, devido a altura do voo, o uso de suas imagens para captura de objetos específicos delas mostra-se um desafio, uma vez que a proporção existente entre objeto e a foto como um todo é muito pequena. Além disso, realizar o processamento dessas imagens em tempo real, de maneira embarcada em um VANT, dada as limitações de hardwares desses equipamentos, mostra-se outro desafio. Disto, entende-se que a primeira motivação do trabalho está em melhorar o desempenho da detecção de objetos considerando os recursos limitados da plataforma de hardware.

Para a tarefa de detecção, como nos demais trabalhos citados nessa dissertação, recorre-se às RNC. Entre elas, a principal diferença está na quantidade de estágios de detecção dos algoritmos, que são aqueles que fazem a detecção em dois estágios, como a Fast R-CNN, e também aqueles que a realizam em apenas um, que é o caso da YOLO(??).

A alteração da arquitetura da RNC influí diretamente no resultado desejado. Contudo, a simples troca de um algoritmo para outro não satisfaz completamente a necessidade de aumentar a acurácia desejada, levando à busca por alterações na estrutura da rede para obter mais interessantes

Assim, o trabalho de (??) propõe um modelo de detecção de objetos baseado em fotografias obtidas por meio de VANTs, chamado UAV-YOLOv8, que se vale da YOLOv8 como rede backbone. No geral, essa nova arquitetura melhora significativamente a precisão e a eficiência da detecção de objetos em imagens capturadas por UAVs, mantendo um baixo consumo de recursos, o que o torna adequado para plataformas de UAVs com capacidades limitadas. Ele é especialmente otimizado para enfrentar os desafios das imagens aéreas, como a detecção de pequenos objetos em fundos complexos.

A oitava versão da YOLO notabiliza-se por ser o estado da arte na ciência de detecção de objetos. Sua estrutura utiliza a arquitetura modificada CSPDarknet53 como backbone, com módulos C2f e SPPF para otimizar a extração e processamento das características de imagem. O neck adota a estrutura PAN-FPN, que melhora a fusão de informações posicionais e semânticas, tornando o modelo mais leve e eficiente. Por fim a head desacoplada usa ramos separados para classificação e regressão de caixas, melhorando a precisão e a robustez da detecção, sem o uso de âncoras (??).

Neste artigo, então, para melhorar sua eficácia nesses cenários, foi proposta uma otimização do modelo utilizando a função de perda WIoU v3, que aprimora a generalização; o mecanismo de atenção BiFormer, que foca em características de alta relevância; e o módulo de processamento de características FFNB, que melhora a fusão de características de diferentes escalas. Essas melhorias resultaram no UAV-YOLOv8, que expande a detecção

de 3 para 5 escalas, aprimorando significativamente a detecção de pequenos objetos.

A experimentação do novo modelo foi proposta de forma a avaliar a precisão de detecção de objetos pequenos e a eficiência de recursos do modelo proposto, UAV-YOLOv8s. Como dataset, o estudo fez uso do dataset VisDrone2019, que inclui fotografias aéreas obtidas por UAVs em várias cidades da China, capturando diferentes cenários e condições de iluminação. Escolheu-se o VisDrone2019 devido à sua riqueza de informações e diversidade. Foi desenvolvido pela Universidade de Tianjin e inclui uma variedade de categorias de objetos de detecção e distribuições variadas de objetos, abrangendo tanto cenários com alta quanto baixa densidade de objetos, e fotos tiradas durante o dia e a noite.

Assim, para a detecção de objetos em fotos capturadas por VANTs, em se tratando de objetos muitos, fundos complexos e recursos de hardware limitados, há um desafio em manter a alta a precisão dos modelos. O UAV-YOLOv8, baseado no YOLOv8, é proposto para otimizar a detecção, introduzindo a função de perda WIoU v3 e o mecanismo de atenção BiFormer, melhorando o foco do modelo e reduzindo a perda de detecção de objetos pequenos. O modelo aprimorado alcança um aumento médio de 7,7% (Figura ??) na precisão sem aumentar seu tamanho, mas a adição de duas camadas de detecção aumenta a complexidade e o tempo de computação. Comparado com outros modelos mainstream, o UAV-YOLOv8 demonstrou desempenho superior, de modo geral (Figura ??). A precisão ainda é baixa para objetos muito pequenos, como bicicletas, e futuras pesquisas seriam necessárias na otimização dessa precisão, podendo elas ficarem na confecção do dataset para melhores resultados, alteração da arquitetura, entre outras abordagens.

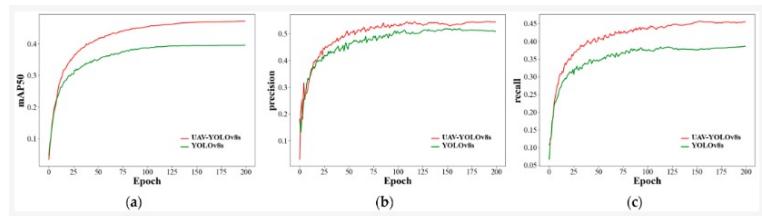


Figura 17 – Comparaçāo da YOLOv8 com a rede modificada

Models	Precision/%	Recall/%	mAP0.5/%	mAP0.5:0.95/%	Model Size/MB	Detection Time/ms	Parameter/ 10^6
YOLOv8n	43.8	33.0	33.3	19.3	6.6	4.2	3.0
YOLOv8s	50.9	38.2	39.3	23.5	22.5	7.7	11.1
YOLOv8m	56.0	42.5	44.6	27.1	49.6	16.6	25.9
YOLOv8l	57.5	44.3	46.5	28.7	83.5	25.6	43.7
Ours	54.4	45.6	47.0	29.2	21.5	19.5	10.3

Figura 18 – Comparaçāo da YOLOv8 com a rede modificada

3.4 Considerações Finais

A Tabela ?? apresenta um resumo de todos os trabalhos relacionados descritos neste capítulo, considerando os seguintes temas:

- *Treinamento utilizando YOLOv8*: Utilização de RNC para realizar treinamento em conjuntos de imagens;
- *Alteração de Parâmetros da RNC para treinamento*: Variação de parâmetros como batch size e otimizadores para treinamento;
- *Utilização de VANT*: Coleta de fotos utilizando VANTS;
- *Subestações de Energia*: Coleta, treinamento e inserção voltada para subestações de energia;
- *Automação de Inserção de RV*: Construção de sistema de inserção automática de objetos em um sistema de RV.

Tabela 1 – Resumo comparativo dos trabalhos relacionados

Trabalhos Relacionados	Alteração de Parâmetros da RNC	Treinamento utilizando YOLOv8	Utilização de VANT	Subestações de Energia	Automação de Inserção de RV
(??)	✓	✓	✗	✗	✗
(??)	✓	✓	✗	✗	✗
(??)	✓	✓	✓	✗	✗
(ZUZA, 2024)	✓	✓	✓	✓	✓

Pela análise da Tabela ??, entende-se que não há ainda um trabalho que reúna todos os tópicos apresentados. É fato que há diversos trabalhos que exploram a detecção de objetos em diversos cenários, utilizando diversas versões da YOLO, inclusive utilizando VANT, para obtenção de imagens. Contudo, no contexto de subestações de energia e identificação de seus equipamentos, há uma lacuna de contribuições. Naturalmente, isso

mostra o quão prolífico esta dissertação se mostra, ao unir todas essas áreas de estudo num propósito ainda não explorado.

No próximo capítulo, serão detalhados os materiais e métodos utilizados para a solução proposta.

4 Materiais e Métodos

4.1 Introdução

Pesquisas relevantes ao tema de treinamento de RNC com imagens aéreas, RV e temas correlatos foram abordados no capítulo anterior, e se tornaram subsídio para compreender o estado da arte dessas temáticas. A partir dessa coletânea de trabalhos, foi traçada a metodologia aqui adotada.

Neste capítulo, portanto, apresenta-se os materiais utilizados nesta pesquisa, assim como os métodos executados para treinamento das imagens coletadas e a abordagem implementada na automação do sistema de RV.

4.2 Dataset

4.2.1 Captura das Fotos

Para o treinamento da rede neural Yolov8, é necessário que sejam fornecidas imagens relevantes do objeto a ser identificado. Neste estudo, foram coletadas fotos aéreas por meio de VANT (modelo DJI Mavic 2, reconhecido por sua capacidade de captura de alta qualidade e manobrabilidade). Todas imagens capturadas foram obtidas no sobrevoo de duas subestações de energia, ambas no Brasil. A primeira subestação é localizada em Porto Velho, no estado de Rondônia, de coordenadas geográficas -8.914705, -63.957887. A segunda, em Araraquara; coordenadas: -21.832444, -48.347566.

No total, foram capturadas 1257 fotos, sendo 548 em Porto Velho e 709 em Araraquara. Todas as imagens foram capturadas em alta resolução e com dimensões variadas, variando entre 4000x3000 até 8000x6000 pixels.

As imagens são capturadas de modo a registrar o máximo de equipamentos possível, cobrindo toda a área das subestações. Todas as fotos são tiradas verticalmente em relação ao solo, garantindo uma visão abrangente e detalhada das instalações das subestações. Devido ao movimento do VANT, os objetos na foto são representados com pequenas variações de angulação em relação ao solo. Em Porto Velho, as fotos foram capturadas no dia 28 de junho de 2023 (Figura ??). Por sua vez, em Araraquara, as capturas foram realizadas dia 13 de junho de 2023 (Figura ??).



Figura 19 – Captura de foto em Porto Velho (à esquerda) e Araraquara (à direita), onde se notam quatro reatores de núcleo de ar na parte inferior.

4.2.2 Seleção das fotos

Conforme o trabalho de (??), o treinamento eficiente com a RNC envolve o fornecimento de imagens relevantes da rede para que haja consistência no caminho desenvolvimento pelo processamento do algoritmo, de forma a conseguir o melhor resultado. Desta forma, foi realizada a seleção das imagens para o treinamento, selecionando apenas aquelas que contém ao menos um objeto de interesse, que neste trabalho, seria o reator de núcleo de ar.

De todo o conjunto de fotos coletadas, apenas 411 fotos foram selecionadas para serem processadas na RNC. Além disso, destas, 60% foram alocadas para o conjunto de treinamento, enquanto 20% para os conjuntos de teste e 20% de validação.

O conjunto de treinamento é utilizado para ajustar os parâmetros internos do modelo. Durante essa fase, a RNC aprende a identificar padrões e características específicas das imagens que correspondem aos objetos de interesse. Naturalmente, objetivando minimizar a função de perda, que mede o quanto bem o modelo está se saindo na tarefa de detecção de objetos.

As fotos usadas para validação, por outro lado, são empregadas para monitorar o desempenho do modelo durante o treinamento. Elas permitem a realização de ajustes nos hiperparâmetros do modelo, como a taxa de aprendizado e o número de camadas. A validação contínua ajuda a prevenir o overfitting, que ocorre quando o modelo se ajusta demasiadamente aos dados de treinamento e perde a capacidade de generalizar para novos dados. Um desempenho consistente no conjunto de validação indica que o modelo está aprendendo de maneira robusta e equilibrada.

Finalmente, o conjunto de teste é utilizado após o treinamento e a validação para avaliar a capacidade de generalização do modelo em dados nunca antes vistos. Este conjunto não influencia o processo de treinamento, mas fornece uma estimativa imparcial do desempenho real do modelo. Um bom desempenho no conjunto de teste sugere que o modelo pode ser confiável para identificar objetos de interesse em situações práticas (??).

Portanto, a divisão dos dados em conjuntos de treinamento, validação e teste é uma prática fundamental em aprendizado profundo, garantindo que o modelo YOLO, ou qualquer outro modelo de detecção de objetos, seja eficiente, robusto e generalizável.

4.2.3 Seleção das fotos

Toda RNC necessita de dados rotulados para aprender a identificar e classificar corretamente os objetos em imagens. Utilizar um Software de marcação, como o LabelImg (??), é essencial nesse processo, pois permite que seja rotulado imagens que irão alimentar a rede, definindo nelas regiões de interesse e associando essas regiões a classes específicas. Nesta dissertação, a classe de interesse trata-se do reator de núcleo de ar. Sem dados rotulados, a RNC não teria a orientação necessária para distinguir diferentes objetos, comprometendo a sua capacidade de realizar previsões.

Para realizar as marcações das fotos selecionadas, foi utilizado o LabelImg, uma vez que trata-se de uma ferramenta de anotação gráfica de código aberto, de fácil instalação e manuseio. Sua interface é intuitiva, pode-se carregar imagens, desenhar caixas delimitadoras ao redor dos objetos de interesse e atribuir rótulos a esses objetos (Figura (??)). As anotações geradas são salvas em um formato compatível, como XML ou TXT, já preparados para serem submetidos ao processamento da YOLO. É essencial que anotações feitas com o LabelImg sejam significativas, pois influenciam diretamente o desempenho do modelo treinado.

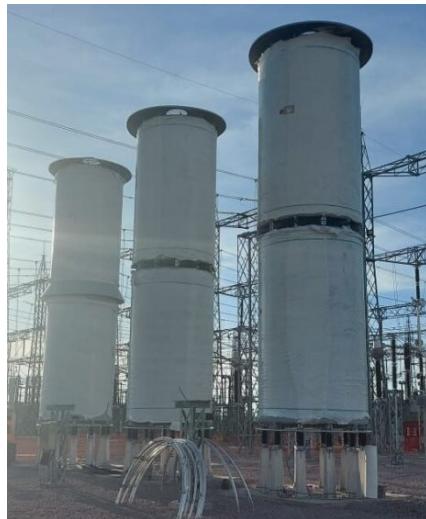


Figura 20 – Reatores de Núcleo de ar em
uma subestação de energia
(??)

4.3 Treinamento

4.4 Considerações Finais

Escrever considerações finais

5 Detalhes da Implementação

5.1 Introdução

Este capítulo irá descrever detalhes da implementação do sistema de treinamento e inserção automática desenvolvida. Para atender as especificações necessárias, uma aplicação foi desenvolvida para facilitar facilitar o treinamento, assim como para automatizar a geração de modelos tridimensionais de fotos inseridas para treinamento.

5.2 Sistema de Inserção

Inserir informações da aplicação

6 Resultados e Discussão

6.1 Introdução

Apresentar os resultados desse trabalho.

6.2 Resultados do Treinamento

Otimizador	Batch Size	metrics/precision(B)	metrics/recall(B)	metrics/mAP50(B)	metrics/mAP50-95(B)	Tempo de Processamento
Adam	2	91.323	90.102	94.022	54.987	33.071h
	4	92.983	91.857	95.4	57.704	36.548h
	8	92.383	91.969	95.006	57.465	31.221h
	16	93.654	92.964	96.148	60.269	29.135h
AdamW	2	93.886	91.954	95.541	58.209	33.114h
	4	93.701	92.102	95.426	57.508	32.744h
	8	93.364	92.215	95.383	57.666	29.209h
	16	95.784	95.379	97.858	68.393	28.013h
SGD	2	94.466	93.062	96.421	61.710	35.448h
	4	94.989	93.779	96.759	65.201	31.018h
	8	95.618	94.621	97.123	66.451	29.787h
	16	96.480	94.718	97.891	68.245	27.633h

Tabela 2 – Resultados dos testes com diferentes otimizadores e tamanhos de batch.

6.2.1 Variação de Parâmetros

Texto da subseção Variação de Parâmetros.

7 Conclusão e Trabalhos Futuros

7.1 Introdução

Neste trabalho, foi proposto um sistema de inserção automática de um equipamento específico de uma subestação de energia em um software de geração de ambientes de RV. Para isso, concorreram a captura de imagens aéreas para ser substrato ao treinamento da RNC; recorreu-se a uma base de modelos digitalizados do equipamento estudado; e foi desenvolvida um sistema automático que vincule essa peça em um ambiente virtual, apenas fornecendo uma imagem em que a mesma esteja representada em uma visão aérea. Esta seção apresentará uma conclusão sobre todo o estudo desenvolvido, assim como trabalhos futuros que poderão partir desse, tanto como melhorias, assim como outras contribuições possíveis à comunidade científica, no devido contexto.

7.2 Conclusão

Conclusão

7.3 Trabalhos Futuros

Antes de tudo, este trabalho propõe-se como o ponto de partida para uma série de futuras implementações e estudos. Aqui, foram treinadas imagens para