UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

MICHELE SOARES DE ANDRADE
Orientador: Eduardo José da Silva Luz

APLICAÇÃO DE DEEP LEARNING NA AVALIAÇÃO DO CONTEÚDO NUTRICIONAL

Ouro Preto, MG
2025

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

MICHELE SOARES DE ANDRADE

APLICAÇÃO DE DEEP LEARNING NA AVALIAÇÃO DO CONTEÚDO NUTRICIONAL

Monografia II apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Eduardo José da Silva Luz

Ouro Preto, MG
2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Michele Soares de Andrade

Aplicação de Deep Learning na Avaliação do Conteúdo Nutricional

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 4 de Abril de 2025.

Membros da banca

Eduardo José da Silva Luz (Orientador) - Doutor - Universidade Federal de Ouro Preto
Daniela Costa Terra (Examinadora) - Mestre - Programa de Pós Graduação em Ciência da Computação (UFOP)
Valéria de Carvalho Santos (Examinadora) - Doutora - Universidade Federal de Ouro Preto

Eduardo José da Silva Luz, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 4/04/2025.

Documento assinado eletronicamente por **Eduardo Jose da Silva Luz**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 08/04/2025, às 20:45, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0886328** e o código CRC **84B2C3F5**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho à Michele do passado, que, mesmo nos seus sonhos mais loucos, não poderia imaginar a força que carregava dentro de si para chegar até aqui.

Agradecimentos

Em primeiro lugar, gostaria de expressar minha profunda gratidão ao meu orientador, Eduardo Luz. Sem ele, este trabalho não seria possível. Muito obrigada por sua dedicação, paciência e por ser o melhor orientador que eu poderia ter.

Ao meu marido, você foi meu porto seguro, minha força nos momentos desafiadores e minha maior motivação. Não há palavras que expressem minha gratidão por tudo o que fez por mim. Esta vitória também é sua.

Ao meu pai Josias (in memorian), você foi o melhor, e jamais vou esquecer seus ensinamentos, obrigada por todo carinho que me deu e o exemplo de homem que foi. Meu amor e saudade por você será eterno.

À minha mãe Marlene, não tenho palavras para agradecer o amor e a atenção que dedicou a mim todos os dias de minha vida. Sem a presença da senhora não seria possível chegar até aqui.

A toda minha família e amigos, que sempre estiveram ao meu lado, celebrando cada avanço e oferecendo apoio nos momentos difíceis.

Aos professores do curso de Ciência da Computação, meu reconhecimento por compartilharem seu conhecimento e por toda a paciência ao longo dessa trajetória. Aos meus colegas de curso, agradeço pela colaboração, pelas risadas e pelo companheirismo nos momentos de desafio e aprendizado.

Às minhas amigas da ONG Código X, que me ajudaram a encontrar um propósito e compartilharam comigo as alegrias e os medos da carreira em tecnologia – vocês são uma inspiração.

Às avaliadoras Valéria e Daniela, meu especial agradecimento por aceitarem participar desta etapa tão importante. É inspirador e significativo ver mulheres ocupando espaços de destaque nesta área.

Por fim, a todos que, de alguma forma, contribuíram para minha formação pessoal e profissional, mesmo que não mencionados aqui, meu muito obrigada.

Resumo

Este trabalho investiga o uso de modelos Vision Transformer (ViT) para estimar macronutrientes presentes em pratos de comida a partir de imagens 2D. A estimativa precisa de macronutrientes é um desafio relevante em aplicações de saúde, especialmente para pessoas com condições como diabetes mellitus tipo 1, que precisam monitorar a ingestão de nutrientes regularmente. O estudo original que serve como referência utilizou um modelo pré-treinado na base JFT-300M, uma coleção de grande escala e diversidade que não está disponível publicamente, dificultando a reprodutibilidade dos resultados. Com o objetivo de avaliar alternativas viáveis, dois modelos ViT foram treinados e avaliados utilizando as bases COYO e ImageNet. A hipótese investigada é que o modelo pré-treinado na base COYO, por ser considerada semelhante à JFT-300M, poderia alcançar resultados superiores ao modelo pré-treinado na ImageNet e mais próximos ao baseline original. Os resultados demonstram que o modelo pré-treinado na JFT-300M apresentou desempenho superior, evidenciado por menores erros médios absolutos (MAE) e percentuais (MAE%) para todas as métricas avaliadas. Por outro lado, o modelo pré-treinado na COYO apresentou os piores resultados, o que sugere que a natureza multimodal e menos curada dessa base limita a extração de características relevantes para a tarefa. O modelo pré-treinado na ImageNet apresentou desempenho intermediário, sendo mais eficaz que o COYO, mas ainda inferior ao baseline. Estes resultados reforçam a importância da curadoria e da relevância do domínio das imagens no pré-treinamento dos modelos. Este estudo contribui ao demonstrar que, embora bases de dados abertas como COYO possam ser alternativas interessantes, a sua eficácia é limitada para tarefas específicas como a estimativa de macronutrientes. Além disso, o trabalho sugere direções futuras, como a expansão da base de dados com pratos típicos brasileiros, a integração de dados 3D e o desenvolvimento de aplicações móveis para democratizar o acesso a ferramentas de análise nutricional.

Palavras-chave: Deep Learning, Vision Transformer (ViT), ResNet-50, InceptionV2, Estimativa Nutricional, Pré-treinamento, JFT-300M, ImageNet, COYO.

Abstract

This work investigates the use of Vision Transformer (ViT) models to estimate macronutrients present in food dishes from 2D images. Accurate macronutrient estimation is a relevant challenge in health applications, especially for individuals with conditions such as type 1 diabetes mellitus, who need to monitor nutrient intake regularly. The original study serving as a reference used a model pre-trained on the JFT-300M dataset, a large-scale and diverse collection that is not publicly available, making it difficult to reproduce the results. To evaluate viable alternatives, two ViT models were trained and evaluated using the COYO and ImageNet datasets. The investigated hypothesis is that the model pre-trained on the COYO dataset, considered to be similar to JFT-300M, could achieve superior results compared to the model pre-trained on ImageNet and closer to the original baseline. The results show that the model pre-trained on JFT-300M achieved superior performance, evidenced by lower mean absolute errors (MAE) and percentage errors (MAE%) for all evaluated metrics. On the other hand, the model pre-trained on COYO showed the worst results, suggesting that the multimodal and less curated nature of this dataset limits the extraction of relevant features for the task. The model pre-trained on ImageNet achieved intermediate performance, proving to be more effective than COYO but still inferior to the baseline. These findings reinforce the importance of dataset curation and domain relevance in model pre-training. This study contributes by demonstrating that, although open datasets like COYO may be interesting alternatives, their effectiveness is limited for specific tasks such as macronutrient estimation. Additionally, the work suggests future directions, such as expanding the dataset with typical Brazilian dishes, integrating 3D data, and developing mobile applications to democratize access to nutritional analysis tools.

Keywords: Deep Learning, Vision Transformer (ViT), ResNet-50, InceptionV2, Nutritional Estimation, Pre-training, JFT-300M, ImageNet, COYO.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

ABNT	Associação Brasileira de Normas Técnicas
DECOM	Departamento de Computação
UFOP	Universidade Federal de Ouro Preto

Sumário

1 Introdução

A estimativa do conteúdo nutricional de pratos de comida a partir de imagens é um desafio significativo na área de visão computacional e nutrição. Com o aumento da preocupação com a saúde e a alimentação balanceada, a capacidade de determinar os valores de calorias, carboidratos, proteínas e gorduras de uma refeição apenas por meio de fotos tornou-se uma área de pesquisa promissora. No entanto, essa tarefa é complexa devido à variabilidade na apresentação dos alimentos, como iluminação, ângulo de captura e disposição dos ingredientes, além da dificuldade em estimar o volume e a quantidade de cada componente apenas com imagens 2D.

Trabalhos recentes, como o de (THAMES et al., 2021), propuseram abordagens para superar esses desafios, utilizando um conjunto de dados grande e variado, aliado a técnicas de aprendizado de máquina. O Nutrition5k, por exemplo, combina imagens 2D com dados de sensores de profundidade para melhorar a precisão das estimativas nutricionais. Outros estudos, como (RUEDE et al., 2021) e (MIN et al., 2020), exploraram a criação de grandes bancos de dados de imagens de alimentos, como o pic2kcal e o ISIA Food-500, que incluem informações detalhadas sobre ingredientes e conteúdo nutricional. No entanto, a maioria dessas abordagens depende de dados adicionais, como sensores 3D ou informações sobre o tamanho das porções, o que limita sua aplicabilidade em cenários do mundo real, onde apenas imagens 2D estão disponíveis.

Um dos desafios enfrentados é em relação à diversidade dos pratos caseiros, que variam amplamente em ingredientes e valores nutricionais, tornando difícil monitorar a dieta de atletas, diabéticos e indivíduos que buscam um controle preciso da ingestão de nutrientes. Para solucionar esse problema, (YAMAKATA et al., 2022) propõem um método de registro de alimentos. Esse método, aliado ao aplicativo RecipeLog [1], permite aos usuários criar novas receitas modificando as existentes, refletindo as receitas reais do dia a dia, contendo informações nutricionais vinculadas a uma lista de ingredientes e um fluxograma de ações de cozimento. Dessa forma, as diferenças entre as receitas reais podem ser facilmente identificadas e registradas, contribuindo para um rastreamento alimentar mais preciso.

Outro desafio abordado é o reconhecimento de categorias de alimentos e ingredientes a partir de fotos de pratos, o que é essencial para a recuperação de receitas e o gerenciamento de registros alimentares. Os trabalhos de (RUEDE et al., 2021) e (CHEN; NGO, 2016) propõem arquiteturas de aprendizado profundo e conjuntos de dados específicos para superar essa dificuldade. Através do uso de redes neurais convolucionais e técnicas de recuperação, os autores demonstram a viabilidade de reconhecer ingredientes e categorias de alimentos em pratos complexos, contribuindo para o avanço das aplicações relacionadas à saúde.

[1] <https://www.recipelog.app/>

Além disso, a estimativa automatizada do conteúdo nutricional das refeições com base em imagens tiradas no mundo real é explorada no estudo apresentado por (MEYERS et al., 2015). Utilizando uma abordagem baseada em aprendizado profundo a qual os autores chamaram de Im2Calories, os pesquisadores desenvolvem um classificador multirrótulo para identificar itens alimentares e uma técnica de segmentação para a classificação refinada e estimativa de volume. Essa abordagem promissora utiliza a câmera móvel como uma ferramenta auxiliar no rastreamento da ingestão de alimentos, superando as limitações de outros métodos.

Neste trabalho, propomos abordar o problema da estimativa nutricional utilizando apenas imagens 2D, motivados pela ampla disponibilidade de câmeras em dispositivos móveis e pela praticidade que essa solução oferece para o usuário final. Apesar dos desafios inerentes à falta de informações de volume e profundidade, acreditamos que técnicas de aprendizado profundo, como redes neurais convolucionais (CNNs) e Vision Transformers (ViTs), podem superar essas limitações, aprendendo padrões complexos diretamente das imagens. Dessa forma, através da análise dos estudos mencionados, espera-se contribuir para o avanço das técnicas de rastreamento alimentar, fornecendo soluções mais precisas, eficientes e acessíveis para indivíduos que buscam um estilo de vida saudável e equilibrado.

1.1 Justificativa

Este trabalho se justifica por sua contribuição ao avanço do aprendizado de máquina aplicado ao reconhecimento nutricional a partir de imagens, por meio da implementação de uma arquitetura promissora para essa tarefa. Além disso, o estudo aborda um problema recorrente na literatura: a dificuldade de replicação de modelos apresentados em artigos científicos, muitas vezes devido ao uso de bases de dados privadas para pré-treinamento. Ao investigar alternativas para contornar essa limitação, este trabalho reforça a importância da reprodutibilidade e da acessibilidade em pesquisas na área.

No estudo inicial da monografia, foram implementados os modelos InceptionV2 e ResNet-50, com base no trabalho que tentamos replicar, que utilizava uma arquitetura derivada do InceptionV2. No entanto, não foi possível alcançar os mesmos resultados obtidos pelos autores. Ao analisar o artigo com mais atenção, percebemos que o desempenho superior do modelo original estava diretamente relacionado ao pré-treinamento em uma base de dados privada denominada JFT-300M. Essa limitação nos motivou a conduzir um estudo mais aprofundado para entender o impacto do pré-treinamento em modelos de reconhecimento de alimentos, utilizando bases de dados públicas e acessíveis.

A autora também possui uma motivação pessoal para a escolha desse tema devido à sua condição de saúde. Sendo portadora de diabetes mellitus tipo 1, ela necessita monitorar de forma precisa a ingestão de carboidratos. Compreender e rastrear o conteúdo nutricional dos alimentos consumidos é fundamental para o controle da glicemia e o manejo adequado da doença.

A partir dessa motivação pessoal, o trabalho visa contribuir para a criação de soluções mais eficazes e acessíveis para pessoas que, como a autora, precisam monitorar de perto sua ingestão de nutrientes.

No âmbito técnico, este trabalho se justifica pela implementação do modelo Vision Transformer (ViT), uma arquitetura de rede neural inovadora que tem demonstrado resultados promissores em tarefas de visão computacional. A dificuldade em replicar os resultados do artigo original, devido à falta de acesso à base de dados pré-treinada utilizada pelos autores, motivou a exploração de duas abordagens distintas: uma com pré-treinamento do ViT na base ImageNet e outra com pré-treinamento na base COYO. Essa investigação não apenas contribui para a compreensão do impacto do pré-treinamento em modelos de reconhecimento de alimentos, mas também oferece insights valiosos para futuras pesquisas na área.

Portanto, este trabalho se justifica pela sua contribuição técnica ao avanço do aprendizado de máquina aplicado ao reconhecimento nutricional, como também pela sua relevância prática e social. Ao implementar o modelo Vision Transformer (ViT) e explorar abordagens para superar a dependência de bases de dados privadas, o estudo promove a reprodutibilidade e acessibilidade na pesquisa científica. Além disso, a motivação pessoal da autora e a proposta deste estudo destaca o potencial do trabalho para impactar positivamente a vida de pessoas que necessitam de monitoramento nutricional.

1.2 Objetivos

O objetivo desta pesquisa é investigar o impacto do pré-treinamento na estimativa de macronutrientes a partir de imagens de pratos de comida, por meio da implementação e avaliação de um modelo Vision Transformer (ViT). Para isso, serão comparadas duas abordagens: uma utilizando um modelo pré-treinado na base ImageNet e outra utilizando o modelo pré treinado na base COYO, buscando compreender as implicações dessa etapa no desempenho do modelo e na reprodutibilidade dos resultados em estudos científicos.

Objetivos específicos:

- Implementar e testar o modelo Vision Transformer (ViT) para o reconhecimento e previsão de conteúdo nutricional (calorias, macronutrientes, etc.) a partir de imagens de pratos de comida;
- Explorar duas abordagens de treinamento:
  - Utilizar pré-treinamento do ViT na base de dados ImageNet.
  - Utilizar pré-treinamento do ViT na base de dados COYO.

- Realizar comparações com os resultados obtidos a partir das duas abordagens entre o artigo base escolhido e as demais implementações utilizando outras arquiteturas que nos levaram a questão de pesquisa;
- Discutir as implicações dos resultados obtidos para futuras pesquisas na área, explorando alternativas para tornar modelos mais acessíveis e aplicáveis a diferentes contextos.

1.3 Hipótese

No trabalho anterior (monografia 1), foi realizada uma tentativa de replicar os resultados do artigo de referência (THAMES et al., 2021), que utilizou uma arquitetura baseada no InceptionV2 para a estimativa de conteúdo nutricional. Apesar de empregar a mesma arquitetura, os resultados alcançados ficaram abaixo do desempenho reportado pelo baseline, que apresentou erros absolutos (MAE) e percentuais (MAE%) significativamente menores. Essa discrepância foi atribuída principalmente ao pré-treinamento do modelo baseline na base JFT-300M, um conjunto de dados proprietário da Google contendo 300 milhões de imagens e 375 milhões de rótulos. A escala massiva e a diversidade dessa base, parecem ter proporcionado representações visuais mais eficazes, algo que não foi possível reproduzir na monografia 1 devido à falta de acesso a essa base.

Diante dessa limitação, surgiu a hipótese para este trabalho: modelos treinados na base COYO, por ser considerada semelhante à JFT-300M em termos de escala e natureza massiva, poderiam alcançar resultados superiores aos obtidos com modelos pré-treinados na ImageNet e, potencialmente, aproximar-se do desempenho do baseline. A COYO, um conjunto de dados extenso e projetado para aprendizado multimodal, foi vista como uma alternativa promissora às bases públicas tradicionais, como a ImageNet, que, embora bem estruturada e amplamente utilizada, possui uma quantidade menor de imagens (cerca de 14 milhões) e um foco em categorias visuais genéricas. A suposta similaridade entre COYO e JFT-300M — ambas caracterizadas por grandes volumes de dados e diversidade — levou à expectativa de que o pré-treinamento na COYO pudesse gerar representações visuais mais precisas, adequadas à tarefa de estimativa nutricional, superando um pré treinamento com a base ImageNet e reduzindo a distância em relação aos resultados do baseline.

Assim, este estudo foi estruturado para testar essa hipótese, comparando o desempenho de modelos Vision Transformer (ViT) pré-treinados nas bases COYO e ImageNet, com o baseline pré-treinado na JFT-300M como referência. A expectativa era que a COYO, por suas características, pudesse ser superior a ImageNet, e também oferecer uma aproximação aos resultados alcançados com a JFT-300M, validando a importância da escala e da diversidade no pré-treinamento para essa tarefa específica.

1.4 Organização do Trabalho

A presente pesquisa encontra-se estruturada da seguinte maneira:

Capítulo 1: Introdução - Neste primeiro capítulo, delineiam-se o contexto do estudo, as motivações subjacentes à sua realização, bem como os objetivos gerais e específicos almejados.

Capítulo 2: Revisão Bibliográfica - No segundo capítulo, procede-se a uma análise dos conceitos essenciais à fundamentação deste estudo, além da analise de trabalhos anteriores relacionados à problemática em questão, os quais desempenharam o papel no aprofundamento das análises empreendidas.

Capítulo 3: Desenvolvimento - No terceiro capítulo, apresenta-se o método adotado juntamente com a métrica empregada na construção deste trabalho. Adicionalmente, expõe-se a base de dados empregada no processo de treinamento, oferecendo detalhamento acerca de sua composição e divisão.

Capítulo 4: Resultados - O quarto capítulo descreve tanto o aparato experimental empregado no estudo quanto os resultados obtidos, seguidos de sua análise.

Capítulo 5: Conclusão - No quinto e último capítulo, são expostas as conclusões advindas deste trabalho de pesquisa, delineando-se também as direções para investigações futuras.

2 Revisão Bibliográfica

2.1 Trabalhos Relacionados

Nesta seção, o objetivo é apresentar uma visão geral de estudos anteriores relacionados ao reconhecimento do conteúdo nutricional a partir de imagens de alimentos. O objetivo é estabelecer uma compreensão abrangente do estado atual do conhecimento e identificar lacunas que exigem mais pesquisas. Pesquisas recentes fizeram avanços significativos na compreensão da nutrição com base em imagens de alimentos e os estudos se concentram em conjuntos de dados, técnicas e resultados associados a essa área de pesquisa.

Vários estudos foram conduzidos para estabelecer bancos de dados para reconhecimento nutricional de alimentos, incluindo aqueles citados em (THAMES et al., 2021),(RUEDE et al., 2021), (MIN et al., 2020), (MIN et al., 2023) e (NARITOMI; YANAI, 2021). No entanto, a criação de tais bancos de dados apresenta um desafio significativo, conforme elucidado por (THAMES et al., 2021). Embora muitos problemas de visão computacional possam usar dados facilmente disponíveis na Internet, a escassez e a imprecisão dos dados de compreensão nutricional dificultam o progresso nesse campo. As imagens de alimentos na Internet são frequentemente apresentadas de forma estilizada em vez de realista, o que dificulta a previsão precisa do conteúdo nutricional a partir de imagens associadas. Na discussão a seguir, vários estudos destinados a construir bancos de dados de conteúdo nutricional serão examinados.

No artigo (RUEDE et al., 2021), é proposto um conjunto de dados, o pic2kcal. O conjunto de dados foi coletado de um site de receitas alemão que inclui listas de ingredientes, instruções de preparo e fotos das refeições. O conjunto de dados também inclui metadados adicionais, como tipo de refeição, avaliação do usuário, tempo de preparação e outras propriedades. Esse esquema resultou na criação do conjunto de dados pic2kcal, composto por 70.000 receitas com 308.000 imagens que contêm informações relacionadas a ingredientes e conteúdo nutricional. Os autores também apontam que o conjunto de dados pic2kcal pode ser usado para resolver outros problemas nas pesquisas sobre reconhecimento nutricional, como detectar refeições veganas ou refeições que se enquadram em dietas específicas, como baixo teor de carboidratos, paleo ou ceto. Apesar de obterem resultados promissores, os autores identificam várias limitações, como a representação limitada do conjunto de dados em relação a outras cozinhas e culturas, a dependência da precisão do banco de dados de nutrientes e a necessidade de saber o tamanho da porção da refeição. Deve-se ressaltar que este conjunto de dados pode ser impreciso, visto que as informações retiradas do site e as fotos correspondentes não são equivalentes, como foi apontado por (THAMES et al., 2021).

Seguindo a mesma ideia de (RUEDE et al., 2021) no qual cria um conjunto de dados a

partir de informações na internet, tem-se o trabalho de (MIN et al., 2020). Os autores apresentam o conjunto de dados ISIA Food-500, bastante abrangente com 500 categorias e quase 400.000 imagens. Os pesquisadores usaram a Wikipedia para construir o sistema conceitual de alimentos e construíram a lista de alimentos de acordo com as "Listas de alimentos por ingrediente" da Wikipedia. Para coletar imagens de alimentos, os pesquisadores usaram um termo de consulta da lista de categorias de alimentos construída e rastrearam imagens candidatas de vários mecanismos de pesquisa (por exemplo, Google, Bing e Baidu).

Pode-se destacar também o conjunto de dados, Food2K. Esse conjunto é proposto por (MIN et al., 2023) no qual possui 2.000 categorias e mais de 1 milhão de imagens. Seguindo a mesma premissa de (RUEDE et al., 2021) e (MIN et al., 2020), para a construção do conjunto foram coletados dados de um site da internet. Os dados coletados foram processados em três fases: (i) Construção de um vocabulário de categorias de alimentos, ou seja, criar uma lista de todos os diferentes tipos de alimentos a serem incluídos no conjunto de dados. (ii) Coleta de imagens de alimentos do conjunto de dados que correspondem às categorias do vocabulário. (iii) Rotulação de cada imagem com sua categoria de alimento correspondente.

Em contrapartida aos trabalhos anteriores, os trabalhos de (THAMES et al., 2021) e (NARITOMI; YANAI, 2021) destacam sobre os desafios associados à estimativa precisa do volume da porção de um determinado alimento. No artigo (THAMES et al., 2021), os autores destacam um problema na área em relação aos conjuntos de dados pois são escassos e muitas vezes imprecisos. Embora os conjuntos de dados encontrados contenham anotações valiosas sobre o prato como, ingredientes e os atributos de preparação, quase sempre carecem de anotações para os tamanhos das porções mostrados nas fotos. Sem anotações precisas sobre o tamanho da porção, aprender a prever o conteúdo nutricional das imagens associadas pode ser difícil e propenso a erros. Dessa forma, os autores apresentam o Nutrition5k, um conjunto de dados contendo 5.000 pratos de comida do mundo real, juntamente com os vídeos correspondentes, imagens de profundidade, pesos dos componentes e anotações de conteúdo nutricional com alta precisão. Os autores exploram a integração de dados do sensor de profundidade para aprimorar as previsões nutricionais. O objetivo principal é treinar um algoritmo de visão computacional que possa prever os valores de calorias e macronutrientes de pratos complexos do mundo real. No entanto, o conjunto de dados tem algumas limitações, como a possibilidade de erro no registro do conteúdo nutricional devido à seleção e exclusão de ingredientes específicos, a restrição geográfica da coleta de dados a uma única cafeteria e a falta de diversidade culinária. Embora o estudo não aborde os desafios de estender o método a conjuntos de dados maiores ou situações além do ambiente do refeitório, ele constitui uma contribuição importante para a compreensão nutricional.

Da mesma forma, os autores (NARITOMI; YANAI, 2021) para a construção de um conjunto de dados empregam um sensor 3D disponível comercialmente, chamado "Structure Sensor", e um aplicativo de digitalização 3D para criar 240 modelos 3D de alimentos e 38

modelos de pratos. Os pratos usados para conter a comida também foram escaneados. Como o mesmo prato foi usado para refeições diferentes, o número de modelos de pratos é menor do que o das refeições. Assim, os autores sugerem uma reconstrução em malha 3D a partir de uma única imagem usando o ResNet, um modelo de Deep Learning. Os pesquisadores consideraram o método eficaz, sendo possível treinar uma rede neural para reconstruir com precisão formas 3D de refeições. No entanto, destacam que pesquisas futuras são necessárias para considerar o tamanho real para obter uma estimativa precisa da ingestão calórica dos alimentos.

Além da base de dados é importante analisar os métodos dos estudos para o reconhecimento dos conteúdos nutricionais das refeições. Em muitos trabalhos é observado a utilização de redes neurais para a previsão dos conteúdos nutricionais como (THAMES et al., 2021), (RUEDE et al., 2021), (YUNUS et al., 2018) e (CHEN et al., 2020).

No trabalho de (THAMES et al., 2021) eles testam o conjunto de dados Nutrition5k em uma arquitetura de rede arquitetura da rede baseada no InceptionV2 com uma resolução de entrada de 256x256 imagens. A rede é otimizada usando o algoritmo RMSprop e pré-treinada usando o JFT-300M. Além disso é utilizado uma abordagem de aprendizado multitarefa. De acordo com os autores, os resultados do aprendizado da rede neural podem prever os valores calóricos e de macronutrientes de um prato complexo do mundo real com maior precisão do que nutricionistas profissionais.

(RUEDE et al., 2021) também propuseram uma abordagem multitarefa baseada na aprendizagem para prever o conteúdo calórico dos alimentos a partir de imagens, utilizando o pic2kcal. Os autores usaram diferentes arquiteturas de redes neurais para estimar as calorias. De acordo com (RUEDE et al., 2021), o conjunto de dados ficou mais rico com informações sobre macronutrientes, o que permitiu a previsão da composição nutricional, além do conteúdo calórico. Os autores sugerem que as propriedades adicionais do conjunto de dados podem ser utilizadas para melhorar os modelos de previsão de calorias, incorporando macronutrientes e ingredientes em configurações multitarefa.

O trabalho de (YUNUS et al., 2018) treina um conjunto de dados para reconhecimento de alimentos usando modelos de rede neural convolucional (CNN) pré-treinados. Os autores utilizam modelos de arquiteturas como, VGG-16, VGG-19, Inception-v3, Inception-v4 e ResNet. Para ajustar os modelos em seu conjunto de dados, os autores removeram a última camada totalmente conectada e anexaram as camadas de dropout, ativações de ReLU e softmax. Foi observado que os modelos baseados no Inception-v3 e no Inception-v4 obteve melhor desempenho do que outros modelos para reconhecimento de alimentos.

Do mesmo modo, em (CHEN et al., 2020) treina um conjunto de dados utilizando uma rede neural convolucional profunda. Os autores realizam experimentos utilizando dois métodos, o modelo de aprendizado multitarefa e de tarefa única. Para os experimentos é proposto quatro arquiteturas derivadas de modelos existentes como VGG-16, ResNet-50, ResNet-101, e SENet-154. Os experimentos fora realizados para uma base de dados de refeições chinesas e demostram

que redes mais profundas podem melhorar o desempenho das tarefas de reconhecimento.

Para avaliar a base de dados, ISIA Food-500, (RUEDE et al., 2021) propõem um método utilizando uma *"stacked global-local attention network"*. A arquitetura consiste em duas sub-redes para reconhecimento de alimentos. Uma sub-rede primeiro utiliza a atenção híbrida do canal espacial para extrair características mais discriminativas e, em seguida, agrega essas características discriminativas em várias escalas de várias camadas na representação em nível global (por exemplo, informações de textura e forma sobre alimentos). A outra sub-rede gera regiões de atenção (por exemplo, regiões relevantes para ingredientes) de diferentes regiões por meio de transformadores espaciais em cascata e agrega ainda mais essas características regionais de várias escalas de diferentes camadas na representação em nível local. Os autores concluem que o método proposto pode servir como uma linha de base confiável para o reconhecimento de alimentos e planejam expandir o conjunto de dados para incluir ainda mais imagens e categorias de alimentos. Também é destacado que o método proposto pode ser adaptado para outras aplicações relacionadas a alimentos, como análise dietética e sistemas de recomendação de alimentos. Bem como, apontam os desafios no reconhecimento de alimentos, como a variabilidade na aparência dos alimentos devido ao estilo de cozimento, iluminação e tamanho da porção. Os autores enfatizam a importância de criar conjuntos de dados maiores e mais abrangentes para enfrentar esses desafios e melhorar a precisão dos modelos de reconhecimento de alimentos.

Em (MIN et al., 2023) é proposto uma *"deep progressive region enhancement network"* para reconhecimento de alimentos utilizando o conjunto de dados Food2K. Este método consiste em dois componentes: o aprendizado progressivo de características locais e aprimoramento de características da região. O método proposto se mostra eficaz em uma variedade de tarefas, incluindo reconhecimento, recuperação, detecção, segmentação e recuperação de receitas multi-modais de alimentos para sua melhor capacidade de generalização. É importante observar, no entanto, que a avaliação do método proposto é limitada ao conjunto de dados sintético Food2K, e seu desempenho em conjuntos de dados do mundo real não foi avaliado. Além disso, o método proposto requer alto poder computacional, o que pode restringir suas aplicações práticas.

Pode-se observar que as metodologias utilizadas na formação dos conjuntos de dados foram consideradas eficazes, embora com certas restrições. Uma dessas limitações diz respeito à representação inadequada do conjunto de dados de práticas culinárias e culturais variadas de várias regiões do mundo. Além disso, a técnica de aquisição de dados e imagens da Internet está sujeita a imprecisões e falha em fornecer uma medição precisa do tamanho das porções, impedindo assim a precisão do reconhecimento de imagens. Consequentemente, esta investigação optou por utilizar a abordagem sugerida na investigação (THAMES et al., 2021) para fabricar um conjunto de dados centrado exclusivamente na gastronomia brasileira. Além disso, para o reconhecimento de refeições, redes neurais serão empregadas para treinar o banco de dados, dada a existência de pesquisas sofisticadas empregando essa metodologia.

2.2 Fundamentação Teórica

2.2.1 Aprendizado de Máquina

Em (GÉRON, 2022) é definido de maneira mais geral que “aprendizado de máquina é uma ciência, bem como arte, de programar computadores para que eles aprendam por meio de um conjunto de dados.” Além disso é fornecido definições de aprendizado de máquina de (SAMUEL, 1959) e (MITCHELL, 1997), que descrevem a natureza implícita e experiencial do aprendizado de computador, respectivamente. A definição de Samuel sugere que a capacidade de aprendizado dos computadores acontece implicitamente, sem programação explícita. Da mesma forma, Mitchell estabelece que um programa de computador aprende por meio da experiência adquirida em relação a uma tarefa específica e a uma medida de desempenho, levando a uma melhoria em seu desempenho à medida que a experiência se acumula. O autor também observa que o acesso a mais dados não garante um melhor desempenho, pois o aprendizado de máquina exige experiências específicas. Um exemplo abordado de aprendizado de máquina é o uso de um filtro de spam, que aprende a identificar e-mails indesejados com base em exemplos previamente classificados. Conforme destacado por (MÜLLER; GUIDO, 2016), a relevância e a aplicabilidade do aprendizado de máquina aumentaram em vários campos, desde recomendações personalizadas em plataformas on-line até aplicações de pesquisa científica.

Deve-se ressaltar que aprender não é memorizar e para isso tem diferentes tipos de aprendizado como (i) Supervisionado (ii) Não Supervisionado e (iii) por Reforço. Para o aprendizado de máquina existem um conjunto de algoritmos e cada um com suas características e usados para diferentes propósitos. Logo, deve-se pensar e analisar qual algoritmo será o ideal para o problema a ser solucionado.

Nas seções subsequentes deste trabalho, os métodos de aprendizado supervisionado e redes neurais serão explorados. Será apresentada uma descrição geral sobre os conceitos relacionados a estes métodos. Com base nesses fundamentos, será possível compreender como as redes neurais podem ser aplicadas de forma eficaz no reconhecimento de conteúdo nutricional em imagens, fornecendo uma base para a análise e interpretação dos resultados obtidos.

2.2.2 Aprendizado Supervisionado

O aprendizado supervisionado é para treinar algoritmos de aprendizado de máquina em que um modelo é treinado para fazer previsões ou tomar decisões com base em um conjunto de dados rotulados. De acordo com (ZHANG et al., 2021) o aprendizado supervisionado é:

> "O aprendizado supervisionado descreve tarefas em que recebemos um conjunto de dados contendo tanto atributos de entrada (características) quanto atributos de saída (rótulos) e tem-se como objetivo produzir um modelo para prever os rótulos com base nas características fornecidas. Cada par de característica-rótulo é chamado

> de exemplo. Às vezes, quando o contexto está claro, podemos usar o termo exemplos para se referir a uma coleção de entradas, mesmo quando os rótulos correspondentes são desconhecidos". (ZHANG et al., 2021), tradução nossa.

O processo de aprendizagem, conforme descrito pelos autores (ZHANG et al., 2021), envolve a coleção inicial de exemplos com características conhecidas, seguida pela seleção de um subconjunto aleatório dessa coleção. Os rótulos verdadeiros são então atribuídos a cada um dos exemplos escolhidos, que podem ser obtidos a partir dos dados existentes. O conjunto de treinamento é formado pela combinação desses dados de entrada e os rótulos correspondentes e inserido em um algoritmo de aprendizado supervisionado. O modelo aprendido resultante pode ser empregado para fazer previsões sobre dados de entrada não vistos anteriormente, permitindo a identificação do rótulo correspondente.

Existem inúmeras abordagens e algoritmos de aprendizado supervisionado disponíveis, incluindo regressão linear, regressão logística, árvores de decisão, máquinas de vetores de suporte (SVM), redes neurais, entre outros. Cada um desses algoritmos tem características específicas e é adequado para diferentes tipos de problemas (GÉRON, 2022).

O aprendizado supervisionado é amplamente usado em várias aplicações, incluindo reconhecimento de padrões, classificação de imagens, diagnóstico médico, detecção de fraudes e muito mais. O nome aprendizado “supervisionado” vem do fato de que o modelo é treinado com exemplos rotulados durante a fase de treinamento, permitindo que ele aprenda a fazer previsões com base nesses rótulos conhecidos (ZHANG et al., 2021).

2.2.3 Regressão

De acordo com (POOLE; MACKWORTH, 2010) “ Regressão linear é um problema de ajustar uma função linear a um conjunto de exemplos de treinamento, nos quais os recursos de entrada e de destino são numéricos.”

No mesmo contexto, (ZHANG et al., 2021) diz que a regressão é um tipo de aprendizado supervisionado que visa prever um valor numérico utilizando um conjunto de características. É frequentemente empregado para estimar o valor de uma casa ou prever os preços futuros das ações. A regressão envolve encontrar um modelo que se aproxime dos valores reais dos rótulos, e muitos problemas práticos podem ser formulados como problemas de regressão. Para fazer isso, vetores de características fixas são frequentemente utilizados, e a minimização da distância entre as previsões e os valores observados geralmente é feita utilizando a função de perda quadrada do erro. (GOODFELLOW; BENGIO; COURVILLE, 2016) também diz que a tarefa de regressão é similar a de classificação, exceto pelo formato de saída diferente.

2.2.4 Classificação

Os problemas de regressão são empregados para a previsão de um valor contínuo, associado a variáveis numéricas. Por outro lado, a classificação é utilizada para tarefas em que se pretende atribuir uma classe ou categoria a uma instância, separando os dados em conjuntos discretos (ZHANG et al., 2021). Nessa perspectiva, a classificação é considerada como uma forma de aprendizado supervisionado, no qual, dada uma entrada, o modelo aprende padrões e características, permitindo que o classificador, ao se deparar com um novo exemplo, possa predizer com a maior precisão possível a sua classe (ZHANG et al., 2021). Existem diversos algoritmos de classificação, tais como: regressão logística, máquinas de suporte vetorial (SVM), árvores de decisão, k-vizinhos mais próximos (KNN), redes neurais, entre outros.

De maneira formal, um classificador é definido por (ZAKI; JR; MEIRA, 2020) como um modelo ou função M que prediz o rótulo de classe y, dado um exemplo de entrada X:

$$y = M(X)$$

onde,

$$\mathbf{X} = (X_1, X_2, \ldots, X_d)^T \in \mathbb{R}$$

é um ponto no espaço d-dimensional e

$$y \in \{c1, c2, \ldots, ck\}$$

é classe predita.

(ZAKI; JR; MEIRA, 2020), em uma abordagem formal, descrevem que:

> "Para construir o modelo de classificação M, precisa-se de um conjunto de treinamento de pontos juntamente com suas classes conhecidas. Diferentes classificadores são obtidos dependendo das suposições usadas para construir o modelo M. Por exemplo, support vector machines utilizam o hiperplano de margem máxima para construir M. Por outro lado, o classificador Bayesiano calcula diretamente a probabilidade posterior $P(c_j|x)$ para cada classe c_j e prevê a classe de x como aquela com a maior probabilidade posterior, $y = argmax_{cj}P(c_j|x)$. Uma vez que o modelo M tenha sido treinado, avalia-se seu desempenho em um conjunto de testes separado de pontos para os quais conhecemos as classes verdadeiras. Finalmente, o modelo pode ser implantado para prever a classe de pontos futuros cuja classe normalmente não conhecemos."(ZAKI; JR; MEIRA, 2020) (tradução nossa)

2.2.5 Redes Neurais

De acordo com (ZHANG et al., 2021) Apos os avanços do poder computacional o cenário da computação em redes neurais foi revolucionado. Devido a essa evolução, alguns pilares atuais

de redes neurais criados há mais tempo foram retomados, como diz (ZHANG et al., 2021).

> "Essa também é uma das razões pelas quais muitos dos pilares do aprendizado profundo, como perceptrons multicamadas (MCCULLOCH; PITTS, 1943), redes neurais convolucionais (LECUN et al., 1998), memória de longo prazo (HOCHREITER; SCHMIDHUBER, 1997) e Q-Learning (WATKINS; DAYAN, 1992) foram essencialmente "redescobertos" na última década, depois de permanecerem relativamente inativos por um tempo considerável."(ZHANG et al., 2021) (tradução nossa)

Uma rede neural é um tipo de modelo computacional que tem a capacidade de aprender padrões e realizar tarefas complexas por meio do processamento de informações. Embora essas redes sejam inspiradas por neurônios cerebrais, elas não simulam neurônios. Em vez disso, eles usam neurônios artificiais conhecidos como unidades, que possuem parâmetros de valor real. De acordo com (POOLE; MACKWORTH, 2010), a estrutura de uma rede neural é composta por várias camadas de neurônios, cada uma com uma função distinta no processo de aprendizagem e tomada de decisão. Os dados iniciais são recebidos pela camada de entrada, que é então propagada pelas camadas intermediárias, chamadas de camadas ocultas. São as camadas ocultas que realizam transformações complexas nos dados, capturando padrões e relacionamentos pertinentes. Em última análise, a camada de saída gera os resultados finais ou as previsões com base no aprendizado adquirido.

Conforme discutido por (POOLE; MACKWORTH, 2010) a tarefa de treinar uma rede neural envolve o ajuste fino dos pesos que conectam os neurônios para otimizar o desempenho da rede em uma tarefa específica. Isso é obtido por meio do uso de algoritmos de aprendizado, como backpropagation. O backpropagation determina a atualização para cada peso com duas passagens pela rede para cada exemplo. Em essência, o algoritmo calcula o gradiente da função de perda em relação aos pesos da rede. Esse gradiente é então usado para atualizar os pesos, melhorando assim o desempenho da rede. Um dos pontos fortes das redes neurais é sua capacidade de aprender e generalizar a partir de uma grande quantidade de dados. Isso é particularmente útil em tarefas como reconhecimento de imagem e processamento de linguagem natural.

Nesse contexto, deep learning é um subconjunto do aprendizado de máquina preocupado com modelos baseados em redes neurais de várias camadas. O termo "profundo" é empregado em referência ao fato de que os modelos são capazes de aprender várias camadas de transformações, conforme explicado no trabalho de (ZHANG et al., 2021). (LECUN; BENGIO; HINTON, 2015) por sua vez, delineiam que uma arquitetura deep learning enfatizando a importância de várias camadas que aprendem e computam mapeamentos não lineares, o que aumenta a seletividade e a invariância da representação. Ao incorporar várias camadas não lineares, os sistemas de aprendizado profundo são capazes de realizar funções complexas e distinguir detalhes finos, ignorando variações irrelevantes, incluindo plano de fundo, pose, iluminação e objetos

ao redor. O Deep Learning compreende uma variedade de arquiteturas e abordagens que são utilizadas para resolver problemas complexos de processamento de dados, incluindo, mas não se limitando a, Redes Neurais Convolucionais (Convolutional Neural Networks - CNNs), Redes Neurais Recorrentes (Recurrent Neural Networks - RNNs), Redes Neurais de Memória de Longo Prazo (Long Short-Term Memory - LSTM) e Gated Recurrent Units (GRU), Redes de Ativação Totalmente Conectada (Fully Connected Feedforward Networks), entre outros.

2.2.6 Vision Transformer (ViT)

O Vision Transformer (ViT) é uma arquitetura que aplica a estrutura de Transformers, originalmente desenvolvida para tarefas de processamento de linguagem natural (NLP), ao domínio da visão computacional. Diferentemente das redes neurais convolucionais (CNNs), que dominam o campo há décadas, o ViT utiliza um mecanismo de auto-atenção para processar imagens, eliminando a necessidade de vieses indutivos específicos para visão, como a equivariância à translação e a localidade, que são intrínsecos às CNNs. Isso permite que o modelo capture relações globais entre diferentes regiões da imagem desde as primeiras camadas, algo que as CNNs só conseguem fazer em camadas mais profundas. Essa abordagem foi proposta por (DOSOVITSKIY et al., 2020) no artigo "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", que demonstrou que os Transformers, quando pré-treinados em grandes volumes de dados, podem superar ou igualar o desempenho das CNNs em tarefas de classificação de imagens.

A arquitetura do ViT é baseada no Transformer original proposto por (VASWANI et al., 2017), com algumas adaptações para lidar com dados visuais. A imagem é dividida em patches de tamanho fixo (por exemplo, 16x16 pixels), que são linearmente projetados em embeddings. Esses embeddings são então concatenados com um token de classificação (semelhante ao token [CLS] do BERT) e embeddings de posição, que preservam a informação espacial dos patches. A sequência resultante é alimentada em um encoder Transformer, que consiste em múltiplas camadas de auto-atenção e redes MLP (Multi-Layer Perceptron). Na figura 2.1 encontra-se uma visão geral do modelo que foi proposto.

Funcionamento do ViT:

- **Divisão da Imagem em Patches:** Dada uma imagem de entrada $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, onde H e W são a altura e largura da imagem, e C é o número de canais (por exemplo, 3 para RGB), o ViT divide a imagem em N *patches* de tamanho $P \times P$. O número de *patches* N é dado por:

$$N = \frac{HW}{P^2}$$

Cada *patch* é então achatado em um vetor de dimensão P^2C, resultando em uma sequência de *patches* $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2C)}$.

- **Projeção Linear dos Patches:** Cada *patch* é projetado em um espaço de dimensão D (dimensão do *embedding*) por meio de uma projeção linear aprendível $\mathbf{E} \in \mathbb{R}^{(P^2C)\times D}$. Essa projeção transforma os *patches* em *embeddings*, que são análogos aos *tokens* em NLP. Além disso, um *token* especial de classificação (`[class] token`) é adicionado ao início da sequência de *patches*. Esse *token* é aprendível e serve como representação global da imagem após a passagem pelo Transformer.

- **Incorporação de Posicionamento (Position Embeddings):** Para preservar a informação espacial dos *patches* na imagem, *embeddings* de posição $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ são adicionados aos *embeddings* dos *patches*. Esses *embeddings* de posição são aprendidos durante o treinamento e codificam a localização de cada *patch* na imagem. Apesar de a imagem ser 2D, o ViT utiliza *embeddings* de posição 1D, que se mostram suficientes para capturar a estrutura espacial da imagem.

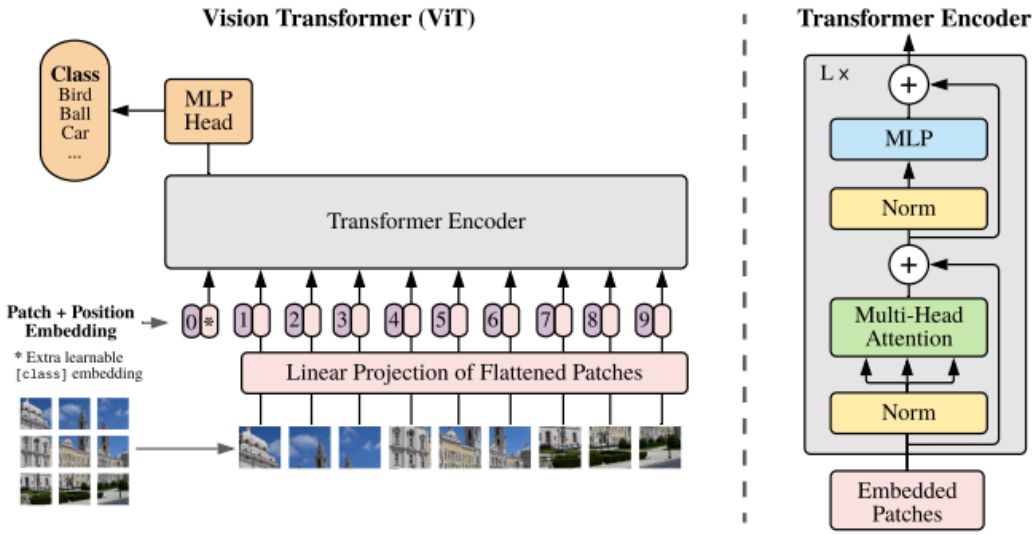
- **Encoder Transformer:** A sequência de *embeddings* (*patches* + *embeddings* de posição) é então processada por um *encoder* Transformer padrão, composto por múltiplas camadas de **Multi-Head Self-Attention (MSA)** e **Multi-Layer Perceptrons (MLPs)**. Cada camada do *encoder* aplica normalização (*LayerNorm*) antes dos blocos de MSA e MLP, e conexões residuais são utilizadas após cada bloco para facilitar o treinamento de redes profundas. A saída do *encoder* é uma sequência de *embeddings*, onde o *embedding* correspondente ao *token* de classificação (`[class] token`) é utilizado como representação da imagem para tarefas de classificação.

- **Cabeça de Classificação:** Para tarefas de classificação, uma cabeça de classificação é adicionada ao final do modelo. Durante o pré-treinamento, essa cabeça é implementada como um MLP com uma camada oculta. Durante o ajuste fino (*fine-tuning*), a cabeça de classificação é substituída por uma única camada linear que mapeia a representação da imagem para as classes do conjunto de dados.

Diferentemente das CNNs, que possuem um viés indutivo intrínseco (como equivariância à translação e localidade), o ViT possui menos viés indutivo específico para imagens. Isso significa que o ViT precisa aprender as relações espaciais e locais a partir dos dados, o que pode ser desafiador em conjuntos de dados menores. No entanto, quando treinado em grandes conjuntos de dados (como ImageNet-21k ou JFT-300M), o ViT supera as CNNs, demonstrando que a escala de dados pode compensar a falta de viés indutivo.

Após o pré-treinamento em grandes conjuntos de dados, o ViT pode ser ajustado para tarefas específicas com conjuntos de dados menores. Durante o ajuste fino, é comum aumentar a resolução da imagem de entrada, mantendo o tamanho dos *patches*. Isso resulta em uma sequência mais longa de *patches*, e os *embeddings* de posição são interpolados para se ajustar à nova resolução.

Figura 2.1 – Visão geral o modelo ViT



Fonte: (DOSOVITSKIY et al., 2020)

O ViT tem sido aplicado com sucesso em tarefas de classificação de imagens, mas seu potencial vai além. A arquitetura pode ser adaptada para outras tarefas de visão computacional, como detecção de objetos e segmentação semântica. Além disso, o ViT abre caminho para a exploração de modelos unificados que combinam visão e linguagem, seguindo a tendência de modelos multimodais como o CLIP e o DALL-E (DOSOVITSKIY et al., 2020).

2.2.7 ResNet-50

A arquitetura ResNet-50 é uma das variantes da família de redes residuais (ResNet), proposta por (HE et al., 2016) no artigo "Deep Residual Learning for Image Recognition". Essa arquitetura foi desenvolvida para resolver o problema de degradação em redes neurais profundas, onde o aumento da profundidade da rede leva a uma saturação e, posteriormente, a uma queda na precisão, mesmo que a capacidade de representação da rede aumente. A ResNet-50, com suas 50 camadas, é uma das versões mais populares da ResNet, equilibrando profundidade e eficiência computacional.

A **ResNet-50** é uma das variantes mais utilizadas dessa família de arquiteturas. Como o nome sugere, ela possui **50 camadas treináveis**, estruturadas em blocos residuais que utilizam *atalhos (skip connections)* para permitir que a informação original da entrada seja preservada ao longo da rede. Essa abordagem facilita a propagação do gradiente, reduzindo problemas como o *vanishing gradient problem*[1] e permitindo o treinamento eficaz de redes mais profundas.

[1] problema em que os gradientes utilizados para atualizar os pesos das camadas anteriores durante o treinamento diminuem exponencialmente à medida que se propagam para trás na rede

A arquitetura da **ResNet-50** é composta por cinco estágios principais, organizados da seguinte maneira:

- **Camada Inicial:**
  - Uma convolução de 7×7 com 64 filtros, seguida de uma camada de *max pooling* com tamanho 3×3 e passo de 2.
- **Blocos Residuais:**
  - A rede utiliza blocos residuais do tipo *"bottleneck"*, que combinam convoluções de 1×1, 3×3 e 1×1, reduzindo a dimensionalidade e melhorando a eficiência computacional.
  - Os blocos são organizados em quatro estágios com **quantidades crescentes de filtros**:
    - **Estágio 1:** 3 blocos residuais, com 64, 64 e 256 filtros.
    - **Estágio 2:** 4 blocos residuais, com 128, 128 e 512 filtros.
    - **Estágio 3:** 6 blocos residuais, com 256, 256 e 1024 filtros.
    - **Estágio 4:** 3 blocos residuais, com 512, 512 e 2048 filtros.
- **Camada Final:**
  - Após os blocos residuais, há uma camada de *average pooling global*, seguida por uma camada *fully connected (FC)* e a ativação *softmax* para classificação.

No artigo, (HE et al., 2016) demonstraram que a ResNet-50 alcançou um erro top-1 de 20.74% e um erro top-5 de 5.25% no conjunto de validação do ImageNet, superando redes anteriores como a VGG-16 e a GoogLeNet. A ResNet-50 também foi usada como parte de um ensemble que alcançou um erro top-5 de 3.57% no conjunto de teste do ImageNet, vencendo a competição ILSVRC 2015.

2.2.8 InceptionV2

A arquitetura Inception-v2, proposta por (SZEGEDY et al., 2016) no artigo "Rethinking the Inception Architecture for Computer Vision", é uma evolução da arquitetura Inception original (GoogLeNet), projetada para melhorar a eficiência computacional e o desempenho em tarefas de visão computacional. A Inception-v2 introduz uma série de otimizações que permitem reduzir o custo computacional e o número de parâmetros, mantendo ou até mesmo melhorando a precisão em benchmarks como o ImageNet.

A Inception-v2 é baseada em princípios de design que visam evitar gargalos de representação, equilibrar a largura e a profundidade da rede, e promover a eficiência computacional. Um dos principais avanços da Inception-v2 é a fatorização de convoluções, onde convoluções com

filtros grandes (5x5 ou 7x7) são substituídas por uma sequência de convoluções menores (3x3 ou 1x7 seguida de 7x1). Essa abordagem reduz o número de operações e parâmetros, mantendo a capacidade de capturar dependências espaciais complexas.

A fatorização de convoluções é uma das principais contribuições da Inception-v2. Convoluções com filtros grandes, como 5x5 ou 7x7, são computacionalmente caras. A Inception-v2 substitui essas convoluções por uma sequência de convoluções menores, como duas convoluções 3x3, que juntas têm um custo computacional menor e mantêm a capacidade de capturar padrões espaciais complexos. Além disso, convoluções assimétricas (1x7 seguida de 7x1) são usadas para reduzir ainda mais o custo computacional, especialmente em camadas intermediárias.

O InceptionV2 mantém a estrutura modular das versões anteriores, organizando suas operações em blocos Inception. A arquitetura segue a seguinte estrutura:

- **Camadas iniciais**:
  - Uma convolução de 3×3 com 32 filtros, seguida de outra convolução 3×3 com 64 filtros, responsável pela extração inicial de características.
- **Blocos Inception aprimorados**:
  - Introdução de módulos Inception com fatoração de convoluções grandes e uso de convoluções assimétricas para melhorar a eficiência.
- **Redução eficiente da dimensão espacial**:
  - Camadas de redução de dimensionalidade utilizando convoluções 1×1 combinadas com convoluções maiores e operações de *pooling*.
- **Camada final**:
  - Um *global average pooling*, seguido de uma camada totalmente conectada (*fully connected*) e ativação *softmax* para classificação.

Os experimentos conduzidos por (SZEGEDY et al., 2016) demonstraram que o InceptionV2 apresenta melhor desempenho e menor custo computacional em comparação com arquiteturas anteriores, como o GoogLeNet e VGGNet. O modelo alcançou redução no número de parâmetros sem perda significativa de acurácia, tornando-se uma alternativa viável para aplicações em visão computacional em larga escala.

3 Materiais e Métodos

Neste capítulo, serão abordados o método empregado no presente estudo, assim como os experimentos conduzidos até o momento e os resultados preliminares obtidos.

3.1 Método

O presente estudo tem como objetivo desenvolver uma abordagem para a identificação do conteúdo nutricional por meio da análise de imagens de pratos de alimentos. O problema é formulado como uma tarefa de regressão, onde uma imagem de um prato de comida serve como entrada, e cinco parâmetros numéricos são extraídos como saída: calorias, massa, proteína, gordura e carboidratos.

O processo de implementação foi conduzido em três fases principais, envolvendo diferentes arquiteturas de redes neurais convolucionais e transformers. Inicialmente, foi realizada a implementação do modelo descrito por (THAMES et al., 2021), que emprega uma arquitetura baseada no InceptionV2. Esse modelo foi escolhido por ser utilizado como referência no trabalho original, sendo importante replicar seus resultados para posterior comparação.

Além disso, para fornecer uma base de comparação e avaliar o desempenho de outras arquiteturas, foi implementado o modelo ResNet-50, conhecido por sua simplicidade e eficácia na extração de características por meio de conexões residuais (HE et al., 2016).

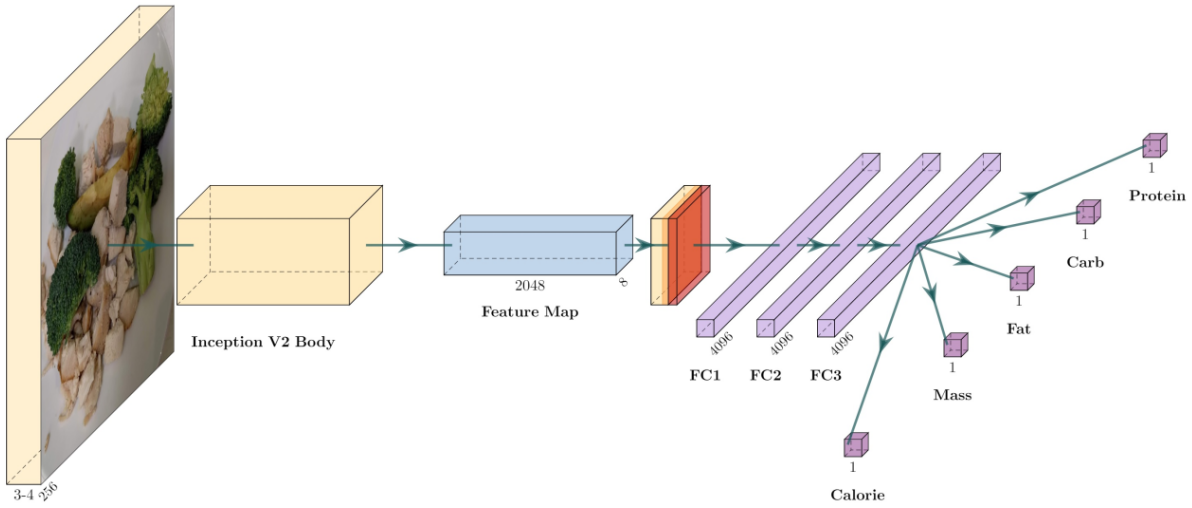
Motivado pela dificuldade de replicar os resultados obtidos por (THAMES et al., 2021), devido ao uso de pré-treinamento com uma base de dados privada (JFT-300M), foi decidido explorar o uso do Vision Transformer (ViT) (DOSOVITSKIY et al., 2020).

A implementação do ViT foi realizada considerando dois cenários distintos:

- ViT pré-treinado na base ImageNet: Um modelo foi inicialmente pré-treinado na base pública ImageNet e posteriormente ajustado (fine-tuning) para a tarefa de estimativa nutricional.
- ViT pré-treinado na base COYO: Outro modelo foi pré-treinado na base COYO, sendo esta uma tentativa de avaliar a influência do uso de diferentes bases de pré-treinamento na tarefa específica de regressão nutricional.

Com base nesse enfoque, o estudo buscou estimar valores relativos às calorias totais, massa, teor de gordura, carboidratos e proteínas em gramas. A proposta abrange a implementação, treinamento e avaliação de diferentes arquiteturas de redes neurais para estimativa nutricional a partir de imagens. A comparação entre os modelos desenvolvidos permite não apenas avaliar a

Figura 3.1 – Visão geral da arquitetura



Fonte: (THAMES et al., 2021)

eficácia do Vision Transformer (ViT), mas também compreender a importância do pré-treinamento e sua influência nos resultados obtidos.

3.1.1 Aprendizagem multitarefa

Os pesquisadores do artigo de referência para este trabalho (THAMES et al., 2021) empregaram uma abordagem de aprendizagem multitarefa , uma técnica utilizada no domínio do Aprendizado de Máquina, em que um único modelo é treinado para executar múltiplas tarefas simultaneamente. Especificamente, para cada tarefa de regressão (isto é, estimativa de calorias, macronutrientes e massa), uma cabeça multitarefa separada é treinada.

A arquitetura da aprendizagem multitarefa se baseou nas saídas da camada "mixed5c"para o caso da InceptionV2, e na saída da penúltima camada, para o caso da ResNet-50. Nesse contexto, um kernel de agrupamento médio [3, 3] com um stride 2 e espaçamento válido foi aplicado. Posteriormente, duas camadas totalmente conectadas (FC) de 4096 dimensões foram introduzidas, compartilhando os parâmetros entre todas as tarefas. Cada tarefa de regressão foi então associada a uma terceira e quarta camada FC finais (também com 4096 dimensões e 1 dimensão, respectivamente). A perda apropriada para cada tarefa foi implementada conforme definida na Equação (3.1).

$$
\begin{aligned}
l_{\text{multi}}(D|W) &= \frac{1}{N} \\
&\quad \sum_{i=1}^{N} \left[l_m(I_i, Y_i^m|W) + l_c(I_i, y_i^{\text{cal}}|W) + l_w(I_i, y_i^w|W) \right] \\
l_m(I, Y^m|W) &= \\
&\quad \frac{1}{|M|} \sum_{j \in M} |\hat{y}_j^m - y_j^m| \\
l_c(I, y^{\text{cal}}|W) &= |\hat{y}^{\text{cal}} - y^{\text{cal}}| \\
l_w(I, y^w|W) &= |\hat{y}^w - y^w|
\end{aligned}
\tag{3.1}
$$

A função de perda geral l_multi é uma combinação ponderada de três funções de perda de subtarefas: perda de regressão de macronutrientes l_m, perda de regressão calórica l_c e perda de peso total l_w. A perda para as regressões citadas usam erro absoluto médio (MAE) como perda de regressão. Os valores dos rótulos previstos para as três subtarefas são representados por $\hat{y}_j^m$, $\hat{y}^{\text{cal}}$, $\hat{y}^m$, enquanto os valores de referência são representados por y_j^m, y^{cal}, y^m.

3.2 Métrica

Através da aplicação dos modelos mencionados anteriormente, o objetivo é estimar parâmetros relacionados às calorias totais, massa, teor de gordura, carboidratos e proteínas, expressos em gramas. Para avaliar a precisão das estimativas de calorias, bem como das massas total e individual de macronutrientes, utilizou-se uma métrica conhecida como Erro Médio Absoluto (MAE, do inglês "Mean Absolute Error"). O MAE é calculado como a média da diferença absoluta entre as previsões e os valores observados, fornecendo um indicador quantitativo da discrepância entre as estimativas e os dados reais. Sua expressão é dada pela seguinte fórmula:

$$
MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.2}
$$

onde:

- n é um o número de amostras.
- y_i é o valor observado ou real para a i-ésima amostra.
- $\hat{y}_i$ é o valor previsto ou estimado para a i-ésima amostra.
- | | representa o valor absoluto.

3.3 Pré treinamento

No presente estudo foi utilizado duas bases como pré treinamento para comparação, a imagenet (RUSSAKOVSKY et al., 2015) e a COYO (BYEON et al., 2022), detalhadas a seguir.

3.3.1 COYO

A base de dados COYO é um extenso conjunto de dados multimodal que contém mais de 700 milhões de pares de imagens e legendas coletados da internet. Desenvolvida e disponibilizada pela Kakao Brain (BYEON et al., 2022), seu principal objetivo é fornecer um recurso de alta escala para o pré-treinamento de modelos de visão computacional e aprendizado multimodal, como modelos de reconhecimento de imagens, classificação e geração de descrições visuais.

A coleta das imagens e legendas foi realizada por meio de técnicas automatizadas de rastreamento na web, seguindo diretrizes éticas e respeitando os direitos autorais dos conteúdos disponíveis publicamente. Cada imagem é associada a uma legenda textual que descreve seu conteúdo, permitindo que os modelos aprendam representações visuais robustas e amplamente generalizáveis.

Como experimento, essa base foi testada em diversos modelos, incluindo o Vision Transformer (ViT). Os modelos foram treinados do zero utilizando o COYO-700M ou seus subconjuntos, alcançando desempenho competitivo em relação aos resultados reportados e às amostras geradas nos artigos originais.

3.3.2 ImageNet

A base de dados **ImageNet** é um dos maiores e mais utilizados conjuntos de dados para o treinamento e avaliação de modelos de visão computacional. Criada como parte do *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*, a ImageNet se tornou importante para tarefas de classificação de imagens e detecção de objetos (RUSSAKOVSKY et al., 2015).

A ImageNet é organizada de acordo com a hierarquia do *WordNet*, uma estrutura semântica que categoriza palavras em sinônimos denominados *synsets*. Cada *synset* é populado com uma média de 650 imagens verificadas manualmente e organizadas por categorias específicas. Ao todo, a ImageNet contém mais de **14 milhões de imagens rotuladas**, distribuídas em mais de **21 mil categorias** (RUSSAKOVSKY et al., 2015).

Para as competições do ILSVRC, é utilizado um subconjunto da ImageNet contendo aproximadamente:

- **1.2 milhões** de imagens para treinamento;
- **50 mil** imagens para validação;
- **100 mil** imagens para teste.

O conjunto de dados de treino é disponibilizado com anotações manuais indicando a presença de categorias específicas em cada imagem. No entanto, as anotações do conjunto de teste são mantidas ocultas, sendo os resultados avaliados por meio de um servidor online.

Esta base desempenha um papel fundamental na área de aprendizado profundo, pois possibilitou o desenvolvimento de arquiteturas que se tornaram referência, como *AlexNet, VGG, ResNet* e *Inception*. Os modelos pré-treinados na ImageNet demonstraram ser altamente eficazes em tarefas de visão computacional, mesmo quando ajustados para outras aplicações específicas por meio de *fine-tuning*.

No presente trabalho, a base **ImageNet** foi utilizada para o pré-treinamento de um dos modelos *Vision Transformer (ViT)*, com o intuito de investigar a influência dessa etapa inicial no desempenho da tarefa específica de estimativa nutricional de pratos alimentares. A ampla diversidade de imagens na ImageNet contribui para que o modelo aprenda representações visuais robustas, melhorando sua capacidade de generalização.

3.4 Base de Dados Nutrition5k

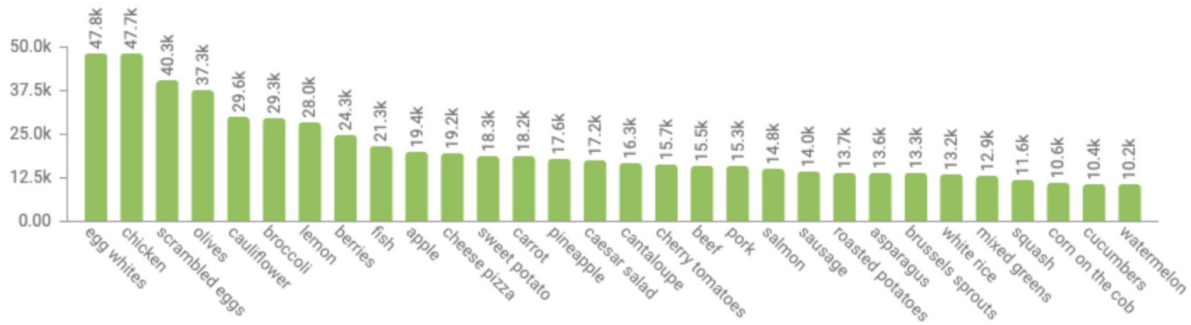
No presente estudo, a obtenção de resultados preliminares foi efetuada por meio da utilização da base de dados denominada Nutrition5k, a qual foi disponibilizada em (THAMES et al., 2021). Os autores desta base propuseram-se a criar um conjunto de dados abrangente, contendo uma variedade representativa de alimentos, enriquecidos com anotações mais precisas de ingredientes. Além disso, a base foi confeccionada com a premissa de alcançar um elevado grau de acurácia, ao mesmo tempo em que se garantisse uma quantidade suficiente de dados aptos para o treinamento de uma rede neural.

Ressalta-se a importância da diversidade de alimentos contemplados na base, uma vez que tal abordagem é fundamental para a capacidade de generalização da culinária no contexto do mundo real. Ademais, é necessário enfatizar que a obtenção de resultados precisos demanda a coleta de dados que proporcionem uma compreensão detalhada do conteúdo presente em cada prato. Em outras palavras, é necessário a incorporação de informações que proporcionem uma percepção da profundidade do prato, incluindo a pesagem dos diferentes componentes do prato.

3.4.1 Descrição e Divisão da Base de Dados

A base de dados, Nutrition5k, é composta por um total de 20.000 vídeos curtos, os quais foram gerados a partir de aproximadamente 5.000 pratos distintos, cada um construído mediante a utilização de mais de 250 ingredientes diversos. Além disso, é válido ressaltar que 3.500 desses pratos estão acompanhados de imagens em formato RGB-D. Cada prato incorpora um rótulo que engloba informações, tais como as quantidades relativas aos alimentos e detalhes referentes aos macronutrientes, cujos cálculos foram derivados de uma fonte de dados secundaria.

Figura 3.2 – Os 30 ingredientes por massa mais comuns



Fonte: (THAMES et al., 2021)

Além disso, a base de dados em questão apresenta uma variedade de tamanhos, porções e quantidades dos alimentos nos pratos, refletindo, assim, na variação observada nos conteúdos relacionados aos macronutrientes. Os autores, conforme documentado em (THAMES et al., 2021), ilustram, na Figura 3.2, os 30 ingredientes de maior recorrência em termos de massa.

No âmbito da estratégia de avaliação, a base de dados foi estruturada em subconjuntos distintos, destinados a treinamento e teste. Nesse sentindo, reservou-se 10% da base para a fase de teste, enquanto que o restante é direcionado ao processo de treinamento.

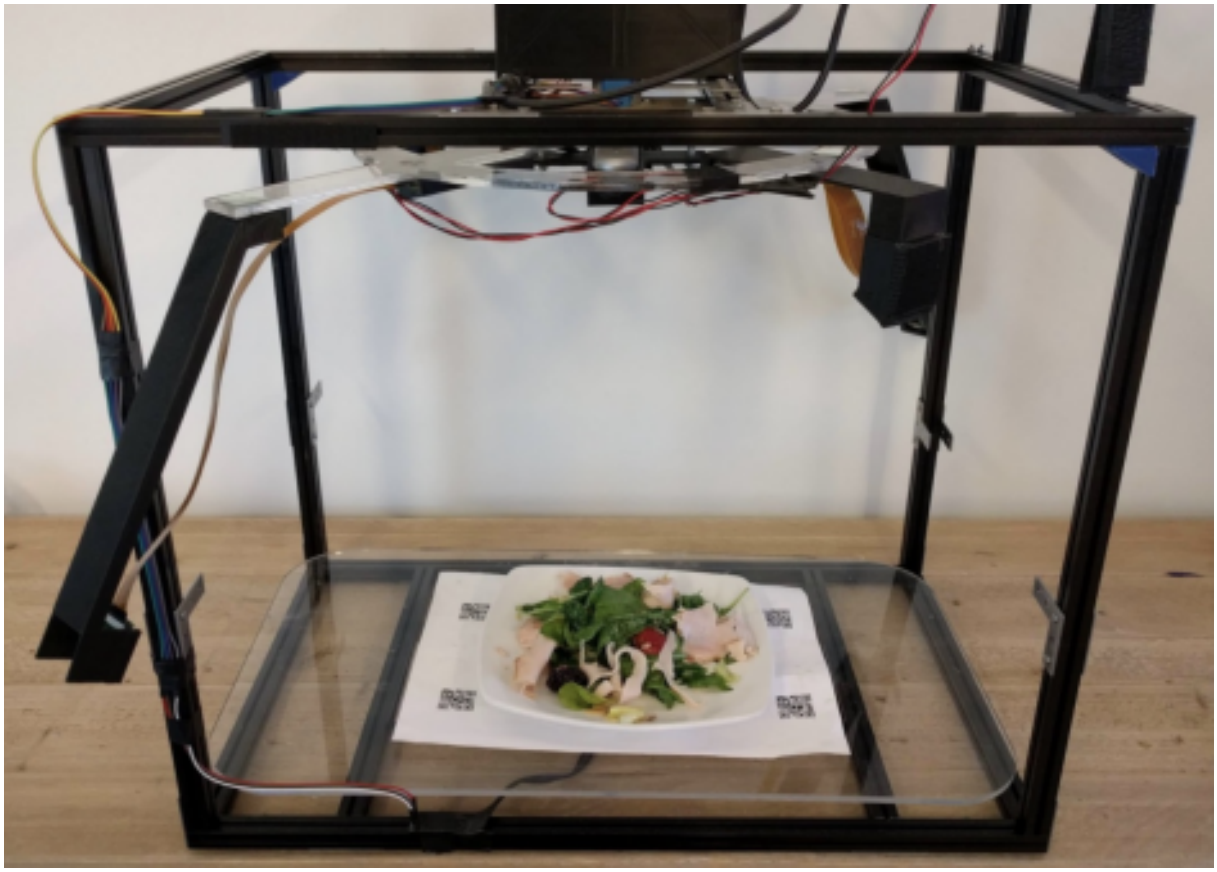
3.4.2 Coleta de Dados

O procedimento de coleta do conjunto de dados Nutrition5k envolveu a utilização de um sistema de sensores customizado, projetado para pesar e escanear individualmente cada prato. Uma abordagem automatizada foi adotada (visualizada na Figura 3.3) para a coordenação e ativação simultânea de todos os sensores, visando reduzir o tempo necessário para cada processo de escaneamento. A coleta dos pratos ocorreu em ambientes de cafeterias do tipo buffet, onde os alimentos eram adicionados individualmente a um recipiente, seguidos por escaneamentos imediatos após cada inclusão.

A determinação do conteúdo nutricional por grama de cada item foi executada através da análise da proporção de ingredientes por peso, utilizando como referência o Banco de Dados de Alimentos e Nutrientes do USDA. Esse valor resultante foi então multiplicado pela medida incremental de peso, à medida que o item era adicionado ao prato, gerando, assim, a anotação básica do conteúdo nutricional.

Durante cada sessão de escaneamento, os seguintes dados foram registrados: a identificação do ingrediente ou receita do item adicionado, quatro gravações de vídeo em formato RGB, uma imagem aérea em formato RGB-D e a medição incremental de peso. As cinco câmeras foram posicionadas em torno e acima do prato, sendo uma voltada diretamente para baixo, enquanto as demais quatro foram distribuídas em cada um dos lados do recipiente. As quatro câmeras de ângulo lateral executaram varreduras simultâneas de 90 graus, capturando, assim,

Figura 3.3 – Equipamento para coleta de dados



Fonte: (THAMES et al., 2021)

uma visualização completa de 360 graus. Estas câmeras alternaram posições a aproximadamente 30 e 60 graus abaixo da linha do horizonte.

As gravações de cada câmera foram realizadas em uma resolução de 1920 × 1080 por um intervalo de aproximadamente 8 segundos por varredura. A imagem de profundidade foi gerada por meio da média dos registros ao longo desse intervalo, com o propósito de eliminar eventuais ruídos e imperfeições na captura. Para a obtenção de dados de profundidade aérea, utilizou-se um sensor Intel RealSense D435, com unidades de profundidade estabelecidas em 1×10^{-4}. Adicionalmente, uma balança digital situada sob o prato registrou o peso incremental, com uma precisão de +/- 1 grama.

3.4.3 Rótulos de Supervisão

A base de dados poder ser representada como $D = \{I_i, Y_i\}_{i=1}^{N}$, sendo que I_i é a imagem, Y_i é o rotulo de supervisão, e N o númeno de exemplos. Para o rotulo de supervisão $Y_i = (y_i^w, Y_i^m, y_i^{cal})$ têm-se: y_i^w rótulo do total do peso; Y_i^m rótulo de macronutrientes; y_i^{cal} rótulo de calorias. Todos os rótulos de supervisão são funções do peso em gramas, onde cada ingrediente K^l no qual l é um índice que abrange todos os ingredientes, da forma:

- y_i^w peso total do prato.
- Y_i^m é um vetor com o peso de cada macronutriente. $M = \{\text{carboidrato}, \text{gordura}, \text{proteína}\}$ para todo $j \in M$, em que $y_{ij} = F_{\text{macro}}(K_{l,j})$, onde F_{macro} é uma função que calcula a quantidade de cada macronutriente.
- $y_i^{cal} = F_{caloria} K_l$ função que calcula total de caloria pelo peso do ingrediente.

4 Resultados

4.1 Experimentos

Neste trabalho, foram implementados e avaliados diferentes modelos de aprendizado profundo para a tarefa de estimativa nutricional a partir de imagens de pratos de alimentos. Inicialmente, a implementação seguiu o trabalho de (THAMES et al., 2021), utilizando a arquitetura **InceptionV2**, também foi implementado para fins de comparação a arquitetura **ResNet-50**. Posteriormente, foram realizados novos experimentos utilizando a arquitetura **Vision Transformer (ViT)**, com diferentes abordagens de pré-treinamento e distintas quantidades de épocas de treinamento.

4.1.1 Configurações Iniciais

Seguindo o protocolo do trabalho original, os modelos **InceptionV2** e **ResNet-50** foram implementados, pré-treinados com a base Imagenet e treinados utilizando imagens de tamanho 256×256 pixels, as quais foram redimensionadas e recortadas no centro para focalizar nas áreas mais relevantes das imagens. A otimização das redes foi realizada com o algoritmo **RMSprop**, utilizando uma taxa de aprendizado inicial de 1×10^{-4}, um momento de $0,9$, um decaimento de $0,9$ e um valor epsilon de $1,0$.

Ressalta-se que as partições entre os subconjuntos de treino e teste foram mantidas constantes em todas as experimentações realizadas. Dessa forma, os mesmos dados foram empregados tanto no processo de treinamento quanto no de teste.

Em relação aos modelos para a análise em duas dimensões (2D), procedeu-se à amostragem de *frames* de vídeos rotacionados. Inicialmente, foi feita uma amostragem utilizando 10 *frames*. Contudo, para obter melhores resultados, optou-se por realizar a amostragem a cada 5 *frames*, como sugerido no trabalho de referência. Sendo assim, a cada cinco *frames* é efetuada a extração de amostras para fins de treinamento.

No trabalho de referência, para os modelos com reconhecimento de profundidade, foi empregado o subconjunto de pratos que dispunham de imagens no formato *RealSense RGB-D*. Em outras palavras, apenas os pratos de comida que possuíam imagens nesse formato foram utilizados no treinamento dos modelos capacitados para a análise de profundidade. Entretanto, no presente trabalho, o interesse é exclusivamente nos modelos RGB (2D), desconsiderando dados com profundidade.

4.1.2 Implementação do Vision Transformer (ViT)

Visando explorar a eficiência de arquiteturas baseadas em Transformers, foram implementados e treinados modelos **Vision Transformer (ViT)** com pré-treinamentos distintos. A implementação do modelo foi realizada utilizando a arquitetura **CLIP ViT** disponibilizada pela biblioteca *Hugging Face Transformers*. O modelo foi adaptado para a tarefa de regressão, com a adição de uma **camada linear de saída** que gera cinco valores correspondentes aos nutrientes: *calorias, massa, gordura, carboidratos e proteínas*.

4.1.2.1 Pré-treinamentos Realizados

Foram considerados dois pré-treinamentos distintos para o modelo ViT:

- **ViT pré-treinado na base ImageNet:** Utilizou-se o modelo pré-treinado na base pública *ImageNet* e posteriormente ajustado (*fine-tuning*) para a tarefa de estimativa nutricional.
- **ViT pré-treinado na base COYO:** Utilizou-se o modelo pré-treinado na base *COYO-700M*, uma base de dados multimodal contendo mais de 700 milhões de pares de imagens e legendas, projetada para aprendizado de representação visual em larga escala.

4.1.2.2 Treinamento e Validação

Cada modelo ViT foi treinado em duas diferentes configurações de quantidade de épocas:

- **60 épocas;**
- **300 épocas.**

A função de perda utilizada foi o **Erro Quadrático Médio (Mean Squared Error - MSE)**, e o treinamento foi realizado utilizando o algoritmo *Adam* com uma taxa de aprendizado ajustada empiricamente. Para melhorar a robustez do modelo, foram aplicadas técnicas de regularização como *Batch Normalization* e *Label Smoothing*.

4.1.2.3 Geração de Predições e Avaliação

Apos as inferências foi calculado as métricas de desempenho, como:

- **Erro Médio Absoluto (MAE);**
- **Erro Médio Absoluto Relativo (%MAE).**

Essas métricas foram calculadas para cada um dos parâmetros nutricionais (*calorias, massa, gordura, carboidratos e proteínas*), permitindo uma análise do desempenho de cada modelo e comparação entre os diferentes pré-treinamentos e quantidades de épocas utilizadas.

4.2 Análise dos resultados

Nesta seção, será apresentada uma análise dos resultados obtidos após a realização dos experimentos. A discussão será dividida em duas partes principais: a comparação entre os modelos ResNet50 e InceptionV2, seguida por uma análise da implementação do modelo ViT.

4.2.1 Resnet50 e InceptionV2

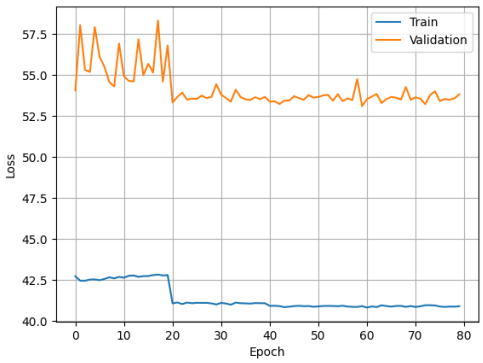
A Tabela 4.1 apresenta uma comparação dos resultados entre os modelos **InceptionV2** e **ResNet-50** treinados com a função de perda L1, e os resultados obtidos a partir do artigo de referência (*Baseline*), que utilizou uma arquitetura baseada no **InceptionV2**. A análise dos valores na tabela revela o desempenho relativo desses modelos na tarefa de estimativa do conteúdo nutricional a partir de imagens de pratos de alimentos.

As colunas "MAE"indicam o erro absoluto médio para cada parâmetro estimado, enquanto as colunas "MAE %"fornecem o erro relativo em termos percentuais. Essas métricas permitem avaliar o desempenho dos modelos em relação à magnitude das estimativas, sendo que menores valores de MAE e MAE % indicam previsões mais precisas.

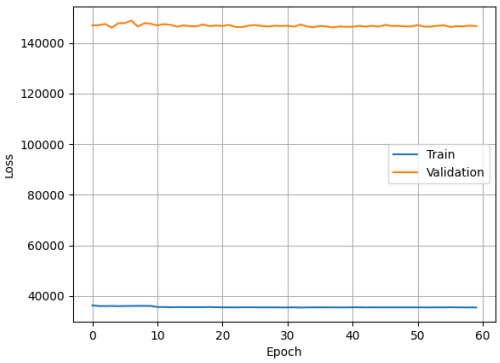
Os resultados evidenciam que o **ResNet-50** apresentou desempenho superior ao **InceptionV2** em todas as métricas avaliadas, apresentando erros absolutos e relativos consistentemente menores. Essa melhoria pode ser atribuída à arquitetura residual do ResNet-50, que facilita o treinamento de redes profundas ao permitir o fluxo eficiente dos gradientes através das conexões de atalho (*skip connections*) (HE et al., 2016). Essas conexões ajudam a mitigar o problema do gradiente desaparecendo, comum em arquiteturas mais profundas como o InceptionV2.

Em relação ao total de calorias, o modelo **InceptionV2** apresentou um **MAE** significativamente maior (**140.88**) em comparação ao **ResNet-50** (**117.38**). Esse padrão é repetido em todas as demais métricas, como massa total, gordura total, carboidratos e proteínas. Vale ressaltar que o desempenho do *Baseline* é superior a ambos os modelos implementados, especialmente para calorias e massa total, indicando que o pré-treinamento utilizado pelo trabalho original contribuiu significativamente para o desempenho final.

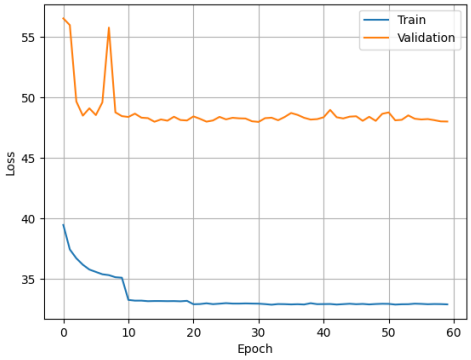
A diferença observada entre o desempenho do *Baseline* e os modelos treinados neste trabalho pode ser explicada pelo uso de um pré-treinamento mais robusto. Os autores do trabalho original utilizaram a base **JFT-300M**, um grande conjunto de dados proprietário contendo mais de 300 milhões de imagens rotuladas, para o pré-treinamento do modelo (SUN et al., 2017). O uso dessa base massiva permitiu que o modelo original aprendesse representações visuais mais robustas e generalizáveis, o que se reflete nos resultados superiores obtidos. É pertinente ressaltar que essa base de dados encontra-se sob propriedade da Google, o que inibe o acesso direto a ela. Essa discrepância reforça a hipótese de que o uso de grandes bases de dados para pré-treinamento, como a **JFT-300M**, é um fator determinante para a obtenção de resultados mais precisos.



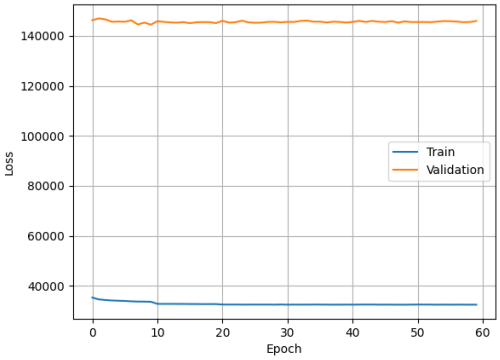
(a) Modelo base ExperimentInception com função de custo L1.



(b) Modelo base Inception com função de custo MSE.



(c) Modelo base ResNet50 com função de custo L1.



(d) Modelo base ResNet50 com função de custo MSE.

Figura 4.1 – Gráficos da custo vs épocas durante o treinamento.

Além disso, os gráficos apresentados na Figura 4.1 ilustram o comportamento do erro ao longo das épocas de treinamento para os modelos **InceptionV2** e **ResNet-50**, considerando as funções de perda L1 e MSE. É possível observar que o **ResNet-50** convergiu de maneira mais estável e consistente, enquanto o **InceptionV2** apresentou uma curva mais irregular, possivelmente indicando maior dificuldade na otimização dos parâmetros.

Esses resultados justificam a escolha de explorar uma abordagem alternativa com o modelo **Vision Transformer (ViT)**, buscando investigar se o pré-treinamento em grandes bases de dados abertas, como a **COYO** e a **ImageNet**, poderia fornecer um desempenho mais próximo ao do modelo pré-treinado na **JFT-300M**.

4.2.2 ViT

Esta seção avalia o desempenho dos modelos Vision Transformer (ViT) pré-treinados nas bases COYO e ImageNet, comparando-os ao modelo baseline pré-treinado na JFT-300M, com o objetivo de responder à questão de pesquisa: o modelo pré-treinado na base COYO, que se assemelha à base JFT-300M utilizada no baseline, se equipara ou supera a base ImageNet? Além

Tabela 4.1 – Comparação entre os modelos Inception (L1-loss), Resnet (L1-loss) e os resultados obtidos no artigo original do dataset.

	Inception Model		Resnet Model		Baseline	
Metric	MAE	MAE %	MAE	MAE %	MAE	MAE %
Total Calories	140.88	55.90	117.38	46.57	70.6	26.1
Total Mass	82.50	43.62	66.83	35.33	40.4	18.8
Total Fat	9.93	73.00	9.02	66.32	5.0	34.2
Total Carb	11.19	58.48	9.80	51.21	6.1	31.9
Total Protein	10.54	67.14	8.92	56.80	5.5	29.5

disso, investiga-se o impacto do pré-treinamento nas bases para a estimativa de conteúdo nutricional a partir de imagens 2D. Conforme (SUN et al., 2017), o volume de dados de pré-treinamento é um fator chave em visão computacional, com desempenho aumentando logaritmicamente em função da escala do dataset, o que fornece um contexto relevante para esta análise. Os resultados, apresentados nas Tabelas 4.2 e 4.3, reportam os erros médios absolutos (MAE) e os erros percentuais (MAE%) em duas configurações de treinamento: 60 e 300 épocas.

Os resultados revelam que o modelo baseline, pré-treinado na JFT-300M, supera os modelos ViT pré-treinados em COYO e ImageNet em todas as métricas analisadas (Total Mass, Total Fat, Total Carb e Total Protein), independentemente do número de épocas. O MAE% do baseline é significativamente menor, destacando sua superioridade para a tarefa de estimativa nutricional. Esse desempenho pode ser atribuído à escala e diversidade da base JFT-300M, 300 milhões de imagens e 375 milhões de rótulos, em média cada imagem tem 1,26 rótulos, proporcionando estimativas melhores.

O modelo pré-treinado na base COYO exibiu o pior desempenho entre os três, com os maiores valores de MAE e MAE% em ambas as configurações (60 e 300 épocas). Por exemplo, com 300 épocas, o MAE para calorias foi de 136,97 (54,34%), contra 95,29 (37,81%) do modelo ImageNet e valores ainda menores do baseline. Esse resultado indica que o pré-treinamento na COYO não é vantajoso para a estimativa de macronutrientes a partir de imagens 2D de pratos de comida. Diferentemente da JFT-300M, que, segundo (SUN et al., 2017), beneficia-se de uma diversidade visual ampla apesar de 20% de ruído nos rótulos, a COYO, embora extensa, pode sofrer com sua natureza multimodal e menos curada, cujo domínio de imagens diverge do contexto específico desta tarefa, limitando a extração de características visuais relevantes.

O modelo pré-treinado na ImageNet apresentou desempenho intermediário, com MAE% inferior ao do COYO, mas superior ao do baseline. Com 300 épocas, o MAE para calorias foi de 95,29 (37,81%), uma melhoria em relação às 60 épocas (87,99, 34,91%), mas ainda distante do baseline. Esse comportamento reflete a capacidade da ImageNet, uma base bem estruturada e rotulada, de fornecer representações visuais úteis para tarefas genéricas, embora não otimizadas para a estimativa nutricional.

Tabela 4.2 – Comparação do Desempenho dos Modelos ViT Pré-treinados nas Bases COYO e ImageNet com o Baseline (60 Épocas)

	COYO		ImageNet		Baseline	
Métrica	MAE	MAE %	MAE	MAE %	MAE	MAE %
Total Calories	144.40	57.29	87.99	34.91	70.6	26.1
Total Mass	91.98	48.63	63.87	33.77	40.4	18.8
Total Fat	10.48	77.04	7.22	53.07	5.0	34.2
Total Carb	12.19	63.68	8.28	43.28	6.1	31.9
Total Protein	11.47	73.03	6.63	42.24	5.5	29.5

O aumento de 60 para 300 épocas resultou em melhorias no desempenho dos modelos ViT. No caso do COYO, o MAE para calorias caiu de 144,40 (57,29%) para 136,97 (54,34%), enquanto no ImageNet oscilou de 87,99 (34,901%) para 95,29 (37,81%). Apesar dessas melhorias, nenhum dos modelos alcançou o desempenho do baseline, sugerindo que o número de épocas, embora relevante, não compensa as limitações impostas pela qualidade e natureza do pré-treinamento.

O gráfico 4.2 complementa a análise quantitativa. O modelo ImageNet apresenta uma curva de treinamento mais baixa e estável, indicando aprendizado eficiente no conjunto de treinamento, mas sua curva de validação é volátil, sugerindo dificuldades de generalização. Já o modelo COYO mostra uma curva de treinamento menos acentuada e valores de perda mais altos, com uma curva de validação também elevada, apontando para um aprendizado insuficiente tanto no treinamento quanto na validação. Esses padrões reforçam que o pré-treinamento na ImageNet é mais eficaz que na COYO, mas ainda inferior ao da JFT-300M.

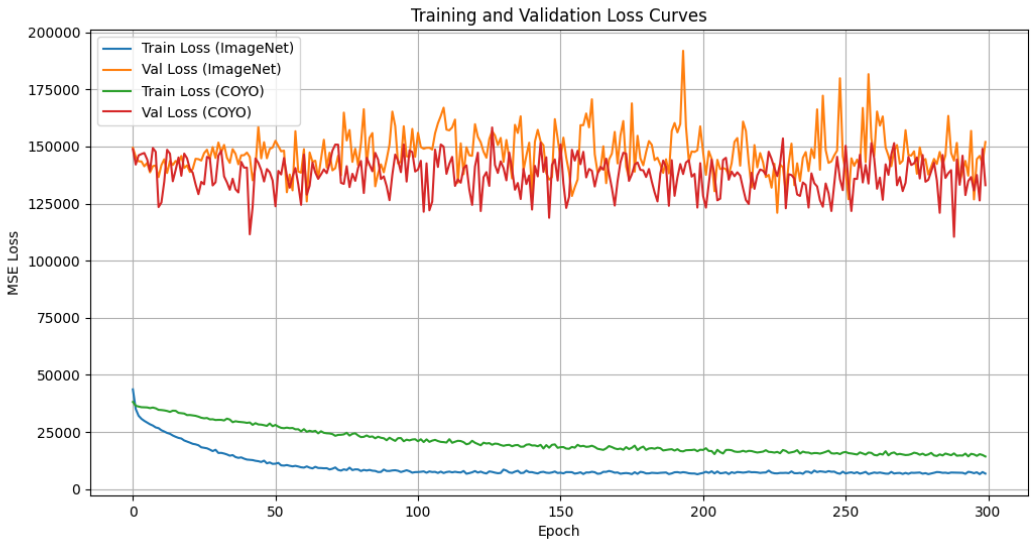
Os resultados demonstram que os modelos pré-treinados na COYO não se saem melhor que o pré-treinamento na ImageNet e nem se equiparam aos resultados do baseline pré-treinado com a base JFT-300M. O impacto do pré-treinamento é claro: bases maiores e mais diversas, como a JFT-300M, geram estimativas melhores, enquanto a qualidade e a relevância do domínio das imagens (como na ImageNet) superam bases grandes, mas menos especializadas (como a COYO). Embora (SUN et al., 2017) mostrem que o desempenho melhora com datasets maiores, nossos achados indicam que, para a tarefa específica de estimativa nutricional em imagens 2D, a relevância do domínio e a curadoria dos dados de pré-treinamento são fatores mais críticos do que o tamanho bruto da base, embora a JFT-300M combine ambos os aspectos para um desempenho ideal.

O código-fonte e os resultados dos experimentos estão disponíveis no repositório GitHub: <https://github.com/michele-andrade/Nutricao-Inteligente>.

Tabela 4.3 – Comparação do Desempenho dos Modelos ViT Pré-treinados nas Bases COYO e ImageNet com o Baseline (300 Épocas)

	COYO		ImageNet		Baseline	
Métrica	MAE	MAE %	MAE	MAE %	MAE	MAE %
Total Calories	136.97	54.34	95.29	37.81	70.6	26.1
Total Mass	82.23	43.48	60.55	32.01	40.4	18.8
Total Fat	10.38	76.27	7.40	54.43	5.0	34.2
Total Carb	11.42	59.64	8.29	43.30	6.1	31.9
Total Protein	10.21	65.05	6.86	43.67	5.5	29.5

Figura 4.2 – Gráfico de curvas de perda de treinamento e validação ao longo de 300 épocas



5 Considerações Finais

5.1 Conclusão

Este trabalho buscou avaliar o impacto do pré-treinamento em diferentes bases de dados na estimativa de conteúdo nutricional a partir de imagens 2D, com foco na comparação entre os modelos ResNet-50, InceptionV2 e Vision Transformer (ViT). A hipótese inicial era de que a base COYO, por sua natureza massiva e supostamente similar à JFT-300M, poderia se equiparar ou superar o desempenho do pré-treinamento na ImageNet para essa tarefa específica. No entanto, os resultados obtidos não corroboraram essa expectativa, refutando a hipótese e revelando limitações significativas do modelo pré-treinado na COYO.

A análise demonstrou que o modelo baseline, pré-treinado na JFT-300M, superou consistentemente todos os outros modelos em termos de erro absoluto médio (MAE) e erro percentual (MAE%), evidenciando sua superioridade na estimativa de métricas como calorias, massa total, gordura, carboidratos e proteínas. Entre os modelos ViT avaliados, o pré-treinado na ImageNet apresentou desempenho intermediário, com erros menores e maior estabilidade em comparação ao ViT pré-treinado na COYO, que exibiu os piores resultados, mesmo após o aumento de 60 para 300 épocas de treinamento. Esse padrão contraria a suposição inicial de que a COYO, por sua escala e características similares às da JFT-300M, poderia oferecer vantagens sobre a ImageNet.

A refutação da hipótese pode ser explicada por diferenças fundamentais entre as bases COYO e JFT-300M, apesar de ambas serem conjuntos de dados extensos. A JFT-300M, um conjunto de dados proprietário da Google contendo 300 milhões de imagens e 375 milhões de rótulos, com uma média de 1,26 rótulos por imagem. Essa escala massiva, combinada com uma curadoria que abrange uma ampla diversidade de domínios visuais, provavelmente permite que os modelos pré-treinados na JFT-300M capturem representações mais robustas e generalizáveis, especialmente para tarefas complexas como a estimativa nutricional. Em contraste, a COYO, embora também seja uma base extensa, possui foco em aprendizado multimodal e carece de uma curadoria tão rigorosa quanto a da ImageNet ou, presumivelmente, da JFT-300M. Essa diferença no domínio das imagens e na qualidade da anotação pode ter limitado a capacidade do modelo ViT-COYO de extrair características visuais relevantes para a identificação de padrões nutricionais em imagens 2D de pratos de comida.

Outro aspecto a considerar é a natureza das tarefas para as quais essas bases foram concebidas. A JFT-300M, desenvolvida para suportar uma ampla gama de problemas visuais em escala industrial, pode conter uma diversidade de categorias e contextos que se alinham mais diretamente com a tarefa de estimativa nutricional, mesmo que indiretamente. Já a COYO, projetada para integrar dados visuais e textuais em um cenário multimodal, pode incluir imagens

cujo conteúdo diverge do foco desta pesquisa, como cenas genéricas ou pouco relacionadas a alimentos. Isso desafia a premissa de que a similaridade em tamanho com a JFT-300M seria suficiente para garantir desempenho comparável, destacando que a qualidade, a relevância do domínio e a curadoria dos dados são fatores igualmente importantes.

Em síntese, a hipótese de que a COYO poderia se equiparar ou superar a ImageNet devido à sua semelhança com a JFT-300M não se confirmou. O pré-treinamento na JFT-300M provou ser o mais eficaz, seguido pela ImageNet, enquanto a COYO apesar de ter apresentado bons resultados não conseguiu superar a ImageNet.

5.2 Trabalhos Futuros

Este estudo oferece um ponto de partida para diversas investigações que outros pesquisadores podem explorar, contribuindo para o avanço da estimativa de conteúdo nutricional a partir de dados visuais. Uma possibilidade seria a ampliação da base de dados atual com a inclusão de uma maior diversidade de pratos típicos brasileiros, capturando a riqueza culinária do país. Esse esforço poderia resultar em um conjunto mais representativo, disponível para a comunidade científica.

Outra direção promissora envolve a experimentação com diferentes parâmetros, técnicas de treinamento e arquiteturas de redes neurais, buscando aproximar os resultados deste trabalho dos alcançados em estudos de referência. Pesquisadores poderiam utilizar a base aqui desenvolvida como base para testar novos modelos ou estratégias de otimização, aprimorando a precisão na estimativa de macronutrientes.

Além disso, uma extensão interessante seria a integração de dados tridimensionais (3D) e o uso de sensores de profundidade, como câmeras RGB-D ou LiDAR, para enriquecer as informações visuais. Estudos futuros poderiam explorar bases de dados que combinem imagens 2D com mapas de profundidade, permitindo a análise de volume e estrutura dos alimentos, o que poderia melhorar a estimativa de porções e densidades nutricionais.

Por fim, pesquisadores poderiam considerar o desenvolvimento de uma aplicação móvel que aproveite os resultados deste estudo, possibilitando que usuários finais — como nutricionistas, chefs ou consumidores — apliquem os modelos em cenários reais de forma prática. Essa iniciativa poderia democratizar o acesso a ferramentas de análise nutricional, ampliando o alcance e a utilidade prática das descobertas.

Esses caminhos sugeridos — desde a expansão da base de dados até a exploração de tecnologias 3D, sensores e soluções aplicáveis — representam oportunidades para a comunidade científica avançar nesta área, combinando inovação metodológica com aplicabilidade prática para enfrentar os desafios da estimativa nutricional.

Referências

BYEON, M.; PARK, B.; KIM, H.; LEE, S.; BAEK, W.; KIM, S. *COYO-700M: Image-Text Pair Dataset*. 2022. <https://github.com/kakaobrain/coyo-dataset>.

CHEN, J.; NGO, C.-W. Deep-based ingredient recognition for cooking recipe retrieval. In: *Proceedings of the 24th ACM international conference on Multimedia*. [S.l.: s.n.], 2016. p. 32–41.

CHEN, J.; ZHU, B.; NGO, C.-W.; CHUA, T.-S.; JIANG, Y.-G. A study of multi-task and region-wise deep learning for food ingredient recognition. *IEEE Transactions on Image Processing*, IEEE, v. 30, p. 1514–1526, 2020.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: "O'Reilly Media, Inc.", 2022.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT press, v. 9, n. 8, p. 1735–1780, 1997.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, p. 115–133, 1943.

MEYERS, A.; JOHNSTON, N.; RATHOD, V.; KORATTIKARA, A.; GORBAN, A.; SILBERMAN, N.; GUADARRAMA, S.; PAPANDREOU, G.; HUANG, J.; MURPHY, K. P. Im2calories: towards an automated mobile vision food diary. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 1233–1241.

MIN, W.; LIU, L.; WANG, Z.; LUO, Z.; WEI, X.; WEI, X.; JIANG, S. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In: *Proceedings of the 28th ACM International Conference on Multimedia*. [S.l.: s.n.], 2020. p. 393–401.

MIN, W.; WANG, Z.; LIU, Y.; LUO, M.; KANG, L.; WEI, X.; WEI, X.; JIANG, S. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2023.

MITCHELL, T. M. *Machine learning*. 1997.

MÜLLER, A. C.; GUIDO, S. *Introduction to machine learning with Python: a guide for data scientists*. [S.l.]: "O'Reilly Media, Inc.", 2016.

NARITOMI, S.; YANAI, K. Hungry networks: 3d mesh reconstruction of a dish and a plate from a single dish image for estimating food volume. In: *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*. [S.l.: s.n.], 2021. p. 1–7.

POOLE, D. L.; MACKWORTH, A. K. *Artificial Intelligence: foundations of computational agents*. [S.l.]: Cambridge University Press, 2010.

RUEDE, R.; HEUSSER, V.; FRANK, L.; ROITBERG, A.; HAURILET, M.; STIEFELHAGEN, R. Multi-task learning for calorie prediction on a novel large-scale recipe dataset enriched with nutritional information. In: IEEE. *2020 25th International Conference on Pattern Recognition (ICPR)*. [S.l.], 2021. p. 4001–4008.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M.; BERG, A. C.; FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, v. 115, n. 3, p. 211–252, 2015.

SAMUEL, A. L. Machine learning. *The Technology Review*, v. 62, n. 1, p. 42–45, 1959.

SUN, C.; SHRIVASTAVA, A.; SINGH, S.; GUPTA, A. Revisiting unreasonable effectiveness of data in deep learning era. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 843–852.

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 2818–2826.

THAMES, Q.; KARPUR, A.; NORRIS, W.; XIA, F.; PANAIT, L.; WEYAND, T.; SIM, J. Nutrition5k: Towards automatic nutritional understanding of generic food. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2021. p. 8903–8911.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

WATKINS, C. J.; DAYAN, P. Q-learning. *Machine learning*, Springer, v. 8, p. 279–292, 1992.

YAMAKATA, Y.; ISHINO, A.; SUNTO, A.; AMANO, S.; AIZAWA, K. Recipe-oriented food logging for nutritional management. In: *Proceedings of the 30th ACM International Conference on Multimedia*. [S.l.: s.n.], 2022. p. 6898–6904.

YUNUS, R.; ARIF, O.; AFZAL, H.; AMJAD, M. F.; ABBAS, H.; BOKHARI, H. N.; HAIDER, S. T.; ZAFAR, N.; NAWAZ, R. A framework to estimate the nutritional value of food in real time using deep learning techniques. *IEEE Access*, IEEE, v. 7, p. 2643–2652, 2018.

ZAKI, M. J.; JR, W. M.; MEIRA, W. *Data mining and machine learning: Fundamental concepts and algorithms*. [S.l.]: Cambridge University Press, 2020.

ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.