UNIVERSIDADE DO ESTADO DO AMAZONAS
ESCOLA SUPERIOR DE TECNOLOGIA
ENGENHARIA ELÉTRICA

TALISSA MOURA AYRES

Classificação de cenas aéreas em sensoriamento remoto: Uma abordagem utilizando dados de imagem e som e *self-supervised learning*.

Manaus
2024

TALISSA MOURA AYRES

Classificação de cenas aéreas em sensoriamento remoto: Uma abordagem utilizando dados de imagem e som e *self-supervised learning*

Projeto de pesquisa desenvolvido durante a disciplina de Trabalho de Conclusão de Curso II e apresentada à banca avaliadora do Curso de Engenharia Elétrica da Escola Superior de Tecnologia da Universidade do Estado do Amazonas, como pré-requisito para obtenção do título de Engenheiro Eletricista.

Orientador: Prof. Dr. Carlos Maurício S. Figueiredo

Manaus
2024

Universidade do Estado do Amazonas – UEA
Escola Superior de Tecnologia - EST

Reitor:
André Luiz Nunes Zodahib
Vice-Reitor:
Kátia do Nascimento Coureiro
Diretora da Escola Superior de Tecnologia:
Jucimar Maia da Silva Júnior
Coordenador do Curso de Engenharia Elétrica:
Jozias Parente de Oliveira, Dr.

Banca Avaliadora composta por: *Data da defesa: <08/02/2024>.*
Prof. Carlos Maurício Serodio Figueiredo, Dr. *(Orientador)*
Prof. Antonio Luiz Alencar Pantoja, Dr.
Prof. Fábio de Souza Cardoso, Dr.

CIP – Catalogação na Publicação

Moura, Talissa

Classificaçao de cenas aéreas em sensoriamento remoto: Uma abordagem utilizando dados de imagem e som e *self-supervised learning* / Talissa Moura Ayres; [orientado por] Carlos Maurício Serodio Figueiredo. – Manaus: 2024. 54 p.: il.

Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica). Universidade do Estado do Amazonas, 2024.

1. Sensoriamento remoto. 2. classificação de cenas aéreas. 3. *deep learning*. 4. *self-supervised learning* 5. modelos multimodais
I. Figueiredo, Carlos Maurício Serodio.

TALISSA MOURA AYRES

CLASSIFICAÇÃO DE CENAS AÉREAS EM SENSORIAMENTO REMOTO: UMA ABORDAGEM UTILIZANDO DADOS DE IMAGEM E SOM E *SELF-SUPERVISED LEARNING*

Pesquisa desenvolvida durante a disciplina de Trabalho de Conclusão de Curso II e apresentada à banca avaliadora do Curso de Engenharia Elétrica da Escola Superior de Tecnologia da Universidade do Estado do Amazonas, como pré-requisito para a obtenção do título de Engenheiro Eletricista.

Nota obtida: 10,0 (dez vírgula zero)

Aprovada em 08/02/2024.

Área de concentração: Inteligência Artificial

BANCA EXAMINADORA

Orientador: Carlos Maurício Serodio Figueiredo, Dr.

Avaliador: Antonio Luiz Alencar Pantoja, Dr.

Avaliador: Fábio de Souza Cardoso, Dr.

Manaus
2024

Dedicatória

Aos meus pais, por todo o amor e suporte em minhas conquistas.

AGRADECIMENTO

Agradeço a minha família, por possibilitar e apoiar minhas conquistas.

Agradeço ao meu orientador, Carlos Maurício, pelo e suporte durante a escrita do trabalho e pelo incentivo durante minha trajetória acadêmica.

Aos meus amigos, por todo o apoio e carinho ao longo dessa jornada.

RESUMO

Realizar classificação de cenas é uma atividade em visão computacional onde modelos conseguem entender um contexto ou ambiente sem focar apenas em classificar um único objeto como acontece em classificação de imagens. Por isso, é uma área de extensa pesquisa, atualmente, por ser utilizada em tarefas importantes como *content based retrieval* e *smart content moderation*. Em adição, quando feita com dados de sensoriamento remoto, ela é importante por auxiliar no entendimento do ambiente ao nosso redor, sendo utilizada em outras tarefas como monitoramento de cidades e classificação do uso da terra. Dando ênfase a classificação de cenas aéreas, muito desses estudos se baseam em utilizar redes neurais convolucionais para essa atividade, então, sendo dependente de uma grande quantidade de anotações para imagens. Por isso, a aplicação de novas técnicas de treino como o *self-supervised learning* (SSL), no qual, se aprende primeiro a gerar representações a partir de *pseudolabels* para depois realizar a tarefa principal, tem sido mais aplicadas na literatura recente. Além disso, a possibilidade do uso de dados multimodais com imagem e som geolocalizados como forma de melhorar o desempenho dos modelos nessa tarefa vem se mostrando uma possibilidade através dos datasets *ADVANCE* e *SoundingEarth*. Sendo assim, o presente trabalho mostra a utilização do SSL e dados audiovisuais de sensoriamento remoto em conjunto com a aplicação dos *vision transformers*, uma nova arquitetura de aprendizado profundo baseada nos mecanismos de *attention*, para geração de representações (*embeddings*). Primeiro, realizou o pre-treino no SoundingEarth, onde se utiliza a *batch triplet loss* para aproximar dados de imagem e som que são pares positivos e afastar pares distintos, em seguida, se aplica as representações em um modelo de regressão logística para classificar as cenas aéreas do *ADVANCE*. Os resultados obtidos foram de precisão, recall e F1-Score acima de 80% para os modelos treinados com os *embeddings* de imagem e som, considerando apenas os *embeddings* de imagem obtemos também resultados acima dos 80% e considerando apenas o áudio obteve-se resultado acima dos 40% para essas métricas.

Palavras chave: Sensoriamento remoto, classificação de cenas aéreas, *deep learning*, *self-supervised learning*, modelos multimodais.

ABSTRACT

Scene classification is an activity in computer vision where models can understand a context or environment without focusing solely on classifying a single object, as in image classification. Therefore, it is an area of extensive research currently, as it is used in important tasks such as content-based retrieval and smart content moderation. Additionally, when performed with remote sensing data, it is crucial for understanding the environment around us, being applied in tasks such as city monitoring and land use classification. Emphasizing the classification of aerial scenes, many of these studies are based on using convolutional neural networks for this activity, thus relying on a large number of annotations for images. Hence, the application of new training techniques such as self-supervised learning (SSL), where the model first learns to generate representations from pseudolabels before performing the main task, has been more widely applied in recent literature. Furthermore, the possibility of using multimodal data with geolocated images and sounds to improve model performance in this task has been demonstrated through the ADVANCE and SoundingEarth datasets. Therefore, this paper demonstrates the use of SSL and audiovisual remote sensing data in conjunction with the application of vision transformers, a new deep learning architecture based on attention mechanisms, for generating embeddings. Firstly, pre-training was conducted on SoundingEarth, using batch triplet loss to bring closer pairs of positive image and sound data and separate distinct pairs. Subsequently, these representations were applied to a logistic regression model to classify aerial scenes from ADVANCE. The results obtained showed precision, recall, and F1-Score above 80% for models trained with both image and sound embeddings. Considering only image embeddings, results were also above 80%, and considering only audio, results were above 40% for these metrics.

Keywords: Remote sensing, aerial scene classification, deep learning, self-supervised learning, multimodal models.

Lista de Figuras

Lista de Tabelas

Lista de Abreviatura

CNN *Convolutional Neural Networks*

FN *False Negative*

FP *False Positive*

GPU *Graphic Processing Units*

ILSVRC *ImageNet Large Scale Visual Recognition Challenge*

ML *Machine Learning*

P Precisão

R Recall

RELU *Rectified Linear Unit*

ResNet *Residual Network*

RoI Região de Interesse

SR Sensoriamento Remoto

SSL *Self-Supervised Learning*

TN *True Negative*

TP *True Positive*

ViT Vision Transformer

SUMÁRIO

INTRODUÇÃO

Durante décadas no campo de visão computacional se utilizou algoritmos tradicionais para resolver problemas clássicos como segmentação/classificação de imagens e detecção de objetos. Por outro lado, com o avanço de recurso computacionais e quantidade de imagens disponíveis se tornou viável o uso de técnicas de *machine learning* aplicados nesse domínio.

Redes neurais pre-treinadas como a *Inception* (SZEGEDY et al., 2014) e *ResNet* (HE et al., 2015) se mostraram poderosas para resolver problemas em classificação de imagens atingindo mais de 70% de acurácia. Através disso, tornou-se possível construir novas aplicações para resolução de diferentes tarefas com imagens como detecção de doenças a partir de imagens médicas (MUKHLIF; Al-Khateeb; MOHAMMED, 2023), reconhecimento de ações (MOUTIK et al., 2023) e também classificação de cenas (ZENG et al., 2021).

Analisando o problema de classificação de cenas temos um desafio interessante para se aplicar modelos de *machine learning*. O modelo precisa aprender quais são as partes semânticas (objetos, texturas e *background*), como eles estão dispostas na cena além de entender qual a relação dessas partes entre si. Apesar de ser um tema de extensa pesquisa, ainda assim existem muitas dificuldades em aplicar tais modelos em situações do mundo real, em específico, classificação de cenas com dados de sensoriamento remoto.

Classificar cenas de sensoriamento remoto tem sido de interesse pois através dessa atividade é possível entender o ambiente ao nosso redor e então poder tomar decisões bem informadas para melhorar o planejamento urbano de cidades (SRIVASTAVA; Vargas-Muñoz; TUIA, 2019), monitorar mudanças no clima (SHAFIQUE et al., 2022) e também auxiliar nas atividades agrícolas como detecção do uso de terra (SINGH et al., 2022).

Dessa forma, pesquisas vem sendo feitas para o desenvolvimento de modelos pre-treinados para classificação de cenas com imagens de sensoriamento remoto, de inicio muitas delas ainda se baseam em *Convolutional Neural Networks* (redes neurais convolucionais - CNNs) para resolver essa tarefa utilizando inclusive modelos pre-treinados com o dataset *ImageNet* (DENG et al., 2009-). Contudo, imagens de sensoriamento remoto se diferenciam muito das existentes na *ImageNet*. As primeiras, contêm em uma imagem mais de um objeto, os quais são distantes entre si e também apresentam classes muito desbalanceadas.

Sendo assim, outras técnicas de treino vem sendo aplicadas para esse tipo de problema. Uma delas é o *self-supervised learning*. Ela se baseia em treinar um modelo com a base de dados em uma *"pretext task"*. A ideia é através de dados não rotulados gerar *pseu-*

dolabels automáticas e utilizar um modelo para aprender representações que distinguiam uma classe da outra. Em seguida, a partir do que foi aprendido com as representações é treinado um novo modelo para a *"target task"*. Os trabalhos de (HEIDLER et al., 2021) e (STOJNIĆ; RISOJEVIĆ, 2021) utilizam dessa técnica e apresentaram bons resultados para a classificação de cenas com dados de sensoriamento. Contudo, trabalhar apenas com dados de imagem mostra uma certa limitação pois imagens são suscetíveis a dois tipos de problemas: grande variação entre imagens de uma classe (*large intraclass variation*) e ambiguidade semântica (*semantic ambiguity*) entre diferentes classes.

Então, uma forma de melhorar a performance desses modelos é trabalhando com dados de diferentes fontes, aprendizado conhecido como *multimodal learning*. Ao combinar dados de imagem e áudio, (HU et al., 2020) mostra que já possível obter bons resultados mesmo utilizando *transfer learning*. Já (HEIDLER et al., 2021) também contribui positivamente a essa afirmação, ao criar uma arquitetura baseada em *self-supervised learning* com as mesmas modalidades de dados e obter uma melhor performance utilizando dados do trabalho anterior, confirmando o potencial uso dessa forma de treinamento.

Considerando o trabalho de (HEIDLER et al., 2021) é visto que ele utiliza as CNNs para aprender as representações do modelo da *pretext task*. Apesar dessas redes obterem bons resultados principalmente por conseguiram extrair informações locais, recentemente, os *vision transformers* ((DOSOVITSKIY et al., 2021),(SCHEIBENREIF et al., 2022)) tem sido de interesse por conseguirem resolver problemas complexos utilizando uma arquitetura mais simples e capaz de aprender características relevantes da imagem a partir do contexto com o mecanismo de *self-attention*.

Sendo assim, o tema deste trabalho está diretamente relacionado ao classificação de cenas com dados de sensoriamento remoto sendo formulado o seguinte objetivo: aplicar o método de *self-supervised learning* utilizando *vision transformers* para classificar cenas aéreas com dados audiovisuais de sensoriamento remoto. Essa proposta pode futuramente ser utilizada para tarefas mais específicas como realizar monitoramento e mapeamento de áreas urbanas ou agrícolas ((JOSHI et al., 2023),(ZHANG et al., 2019)) assim como auxiliar em tarefas como realizar detecção de objetos para imagens de satélite ou drone. ((BYUN et al., 2021),(OSCO et al., 2021))

1 PROBLEMA DE PESQUISA

Dados de sensoriamento remoto são importantes pois através deles entendemos o que acontece na superfície da Terra. É possível realizar planejamento urbano de cidades, auxiliar na gestão de recursos naturais assim como gerenciar plantações de terras agrícolas e entre outras atividades. Técnicas de *machine learning* e *deep learning* tem sido aplicadas e ajudado setores importantes na tomada de decisões. Contudo, o desenvolvimento de soluções para área ainda é muito dependente de dados rotulados que precisam de uma grande quantidade de tempo e recurso financeiro para serem construídos.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Construir um modelo que aprenda a gerar representações de dados de imagem e som de sensoriamento remoto com a abordagem de *self-supervised learning* utilizando *vision transformers* e avaliar sua eficácia em classificar cenas aéreas audiovisuais de sensoriamento remoto.

2.2 OBJETIVOS ESPECÍFICOS

- Fazer levantamento de datasets de imagens e áudio geo-localizados, assegurando sua preparação adequada para análise.
- Implementar a abordagem de *self-supervised learning* para realizar o pré-treino com audiovisuais de sensoriamento remoto não rotulados como forma de aprender a criar representações com som e imagem para serem utilizadas em outras tarefas que necessitam desses tipos de dados, além de trabalhar com uma arquitetura baseada em *vision transformers*.
- Avaliar os resultados obtidos pelo modelo aprendido utilizando métricas como acurácia, *precision* e *recall* em dados rotulados para examinar a eficácia da abordagem com pré-treino utilizando *self-supervised learning* com *vision transformers* em dados não rotulados.

3 JUSTIFICATIVA

O problema de classificação de cena tem relevância pois a partir da análise do ambiente é possível tomar decisões importantes em relação ao planejamento urbano de cidades como monitorar desastres ambientais e analisar o uso de terra além de poder auxiliar atividades econômicas como na agricultura para verificar surgimento de pragas e verificar a saúde da terra agrícola por exemplo. Atualmente, métodos utilizando deep learning aproveitam da grande quantidade de dados de sensoriamento remoto disponíveis para aplicar

técnicas de treinamento supervisionado baseado em redes neurais convolucionais (CNNs) para realizar essa tarefa.

Para executar treinos supervisionados é necessário ter uma grande quantidade de dados rotulados disponíveis, porém, isso é um processo lento e custoso para ser realizado. Dessa forma, esse trabalho se justifica ao propor uma alternativa promissora, a utilização do *self-supervised learning* permite que modelos de classificação eficazes para cenas aéreas sejam construídos a partir de dados de sensoriamento remoto que não necessitam de anotações prévias. Ao empregar técnicas inovadoras como *vision transformers* e redes neurais convolucionais, esta abordagem visa alcançar, ou até mesmo superar, a acurácia dos métodos tradicionais.

Além disso, o tema também engloba diversos conceitos estudados ao longo do curso de engenharia elétrica, tais como:

- Álgebra Linear I e II;
- Cálculo I, II, III e IV;
- Probabilidade e Estatística;
- Linguagem de Programação I;
- Processamento Digital de Imagens;
- Padrões de Compressão de Áudio e Vídeo;
- Tópicos Especiais para Computação II.

Em síntese, esta pesquisa não apenas busca propor uma abordagem alternativa por meio do *self-supervised learning*, mas também aspira a alcançar, ou até mesmo superar, a acurácia dos métodos tradicionais de classificação de cenas aéreas. Ao realizar uma análise comparativa abrangente, espera-se contribuir substancialmente para o avanço da pesquisa em machine learning, enquanto oferece soluções práticas para aprimorar a análise de cenas em contextos do mundo real.

4 REFERENCIAL TEÓRICO

4.1 Visão computacional

A visão computacional é um dos ramos da inteligência artificial que procura desenvolver algoritmos e modelos para coletar, processar e analisar dados de imagens ou vídeos. Seu objetivo é conseguir extrair informações relevantes e fazer decisões baseadas no que é ”visto”. Tarefas clássicas nesse domínio, como detecção de bordas, estimação de movimento e segmentação semântica envolvem a aplicação dessas técnicas. Com a evolução dos recursos computacionais e a crescente adoção do aprendizado profundo (*deep learning*), a

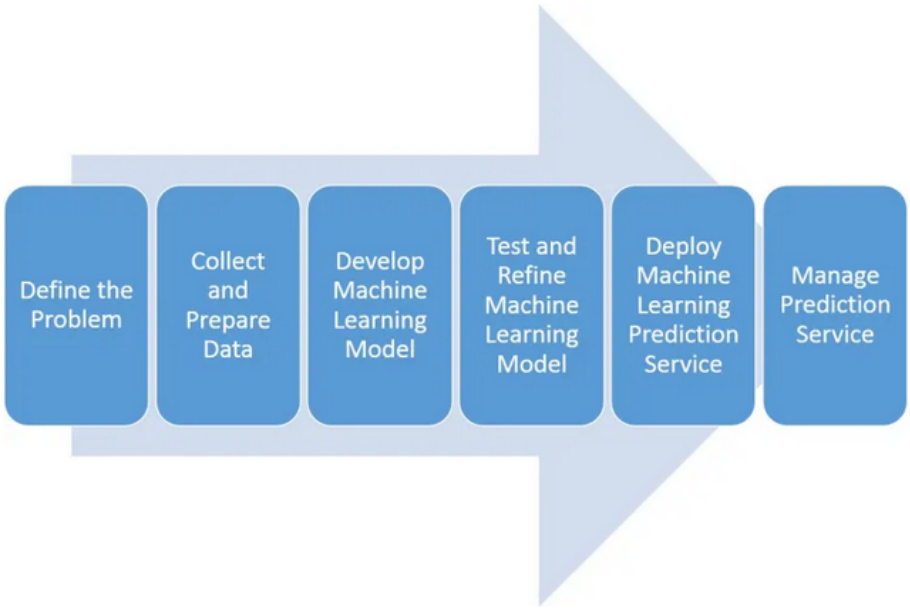
área está experimentando um interesse substancial e alcançando avanços significativos na resolução de suas atividades.

Modelos de *machine learning* são conhecidos por conseguirem extrair e aprender padrões para a resolução de problemas. A grande vantagem deles é saber lidar com grandes quantidades de dados e serem fáceis de manter pois uma vez construído basta obter novos dados para retreinar o modelo. Para tarefas de visão computacional tais características permite ser possível solucionar muitas atividades complexas e também construir aplicações que utilizem desses recursos como sistemas de recomendação de vídeos ((LI et al., 2017)), *face authentication* ((ZULFIQAR et al., 2019)) e *optical character recognition* ((MEMON et al., 2020)) em documentos.

4.2 *Machine Learning*

Machine learning (ML) é um dos ramos da inteligência artificial interessado em desenvolver algoritmos que através dos dados conseguem aprender padrões e fazer previsões de maneira independente sem ser explicitamente programados para isso. Algumas dessas técnicas são baseadas em estatística como Regressão Logistica e Regressão Linear, já outros tem como base estrutura de dados mais tradicionais como árvores de decisão

Figura 1 – Passos para resoluções de problemas com ML.



Fonte: Rodriguez (2023)

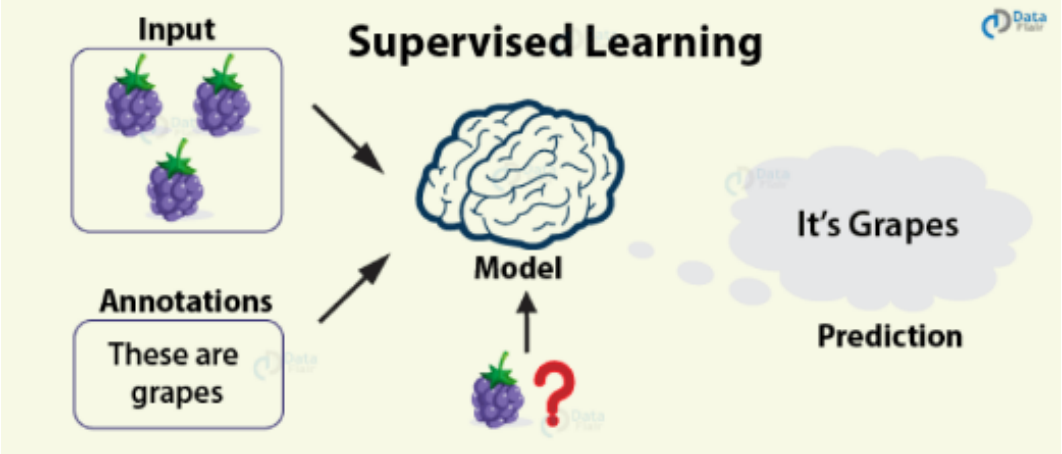
No mundo atual, com a crescente quantidade de dados disponíveis algoritmos de ML tem sido muito presentes em nossa realidade. Algumas das suas aplicações envolvem classificação de imagem ((LIU; TIAN; XU, 2019),(MUKHLIF; Al-Khateeb; MOHAMMED, 2023),(ALZUBAIDI et al., 2021)) e textos ((QIANG et al., 2022)), modelos de

reconhecimento de fala ((RAVENSCROFT et al., 2021)) assim como problemas mais complexos como previsão de risco de crédito ((BHATORE; MOHAN; REDDY, 2020)).

Para resolver tarefas utilizando ML é necessário entender os tipos de treinamento existentes desses algoritmos. Cada um é específico dos tipos de dados existentes e qual o tipo de problema solucionar. Eles são os listados abaixo:

- Supervisionado ou *supervised*: O treinamento supervisionado é baseado em rótulos (ou *labels*) que definem o que o modelo precisa prever. Se é necessário prever classes, esse tipo de problema é conhecido como classificação, caso o valor seja numérico, então, se trata de um problema de regressão.

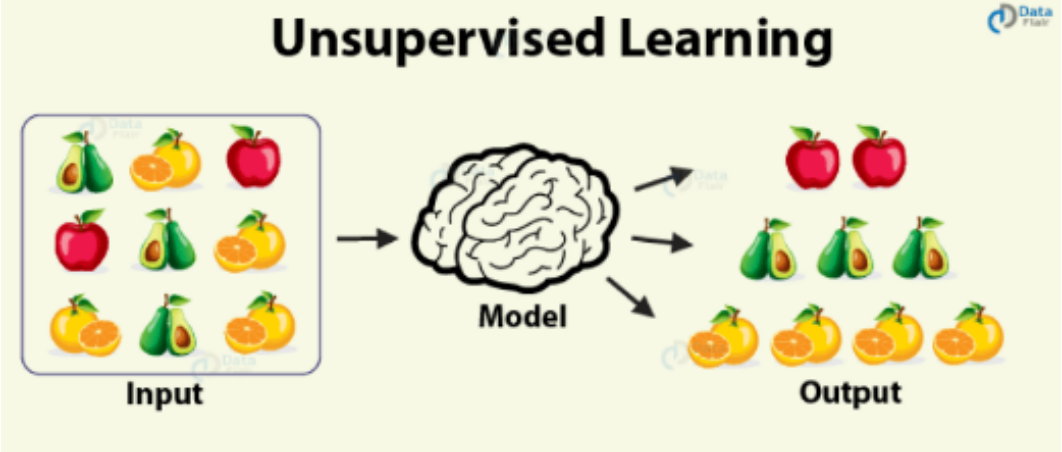
Figura 2 – Treino com *Supervised learning*



Fonte: Shah (2023)

- Não-supervisionado ou *unsupervised*: Nesse caso, o modelo aprende padrões dos próprios dados sem eles terem labels. Se busca nesse caso resolver alguns desses problemas:

Figura 3 – Treino com *Unsupervised learning*.

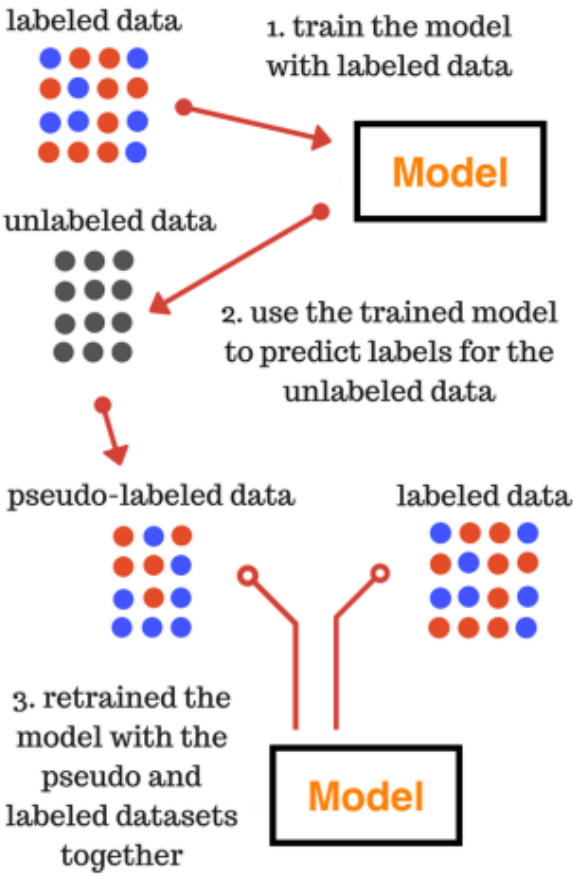


Fonte: Shah (2023)

  - Encontrar grupos similares em um conjunto de dados, técnica conhecida com *clustering*
  - Definir a distribuição da população, técnica conhecida como *kernel density estimation*
  - Reduzir dados multi-dimensionais para dimensões menores com o objetivo de visualização com algoritmos como *Principal Component Analysis*.

- Semi-supervisionado ou *semi-supervised*: Nem sempre é possível obter labels para todos os dados, então por isso existe a abordagem semi-supervised, nela se treina um modelo a partir de dados com labels para gerar *pseudolabels* (”labels estimadas”) aos dados sem labels. Em seguida, com o acréscimo desses novos dados se retreina o modelo para gerar previsões.

Figura 4 – Treino com *semi-supervised learning*.



Fonte: Teksands (2021)

Para construção de modelos com machine learning, na primeira parte, se divide os dados para dados de treino e teste e então é definido dois momentos:

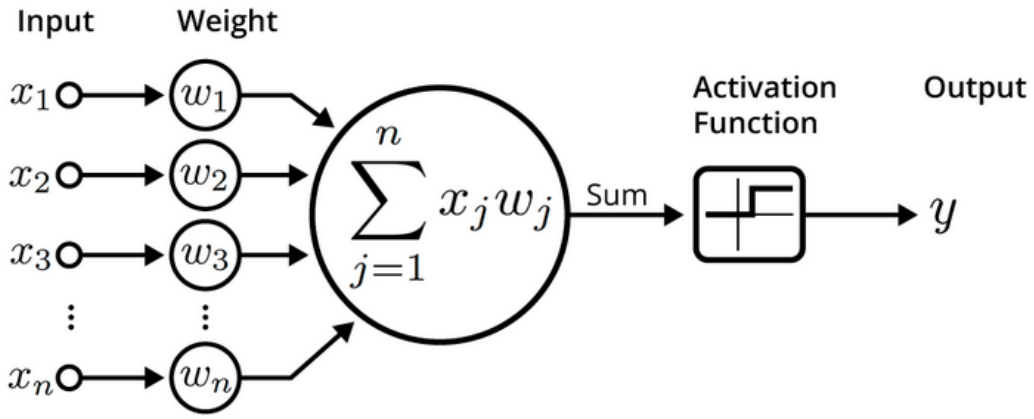
- Treino: onde o modelo utiliza os dados de treino para aprender padrões.
- Teste: com o modelo treinado, ele é testado com os dados de teste e se mede o quão bem ele consegue resolver uma determinada tarefa. Métricas mais conhecidas para problemas de classificação é acurácia, *recall* e *precision*, já para regressão existe a *mean square error* e *root mean square error*.

4.3 *Deep Learning*

Para conceituar *deep learning*, primeiro, é necessário comentar sobre redes neurais artificiais. Essas são algoritmos que emulam o funcionamento do nosso cérebro com neurônios e sinapses. Elas são compostas por neurônios artificias que são pequenos nós os quais através dos dados de entrada (*inputs*) geram um ”sinal”para outros neurônios. Esse sinal é um valor numérico gerado a partir de uma função não linear que soma os valores dos dados de entrada. A ligação entre dois neurônios tem o nome de vértice e possui um

peso (outro valor numérico que define a força dessa conexão). Um neurônio pode possuir um ”valor de limiar”(*threshold*) e enviar um ”sinal”ao próximo neurônio apenas quando a soma dos dos sinais agregados recebidos passarem de um valor.

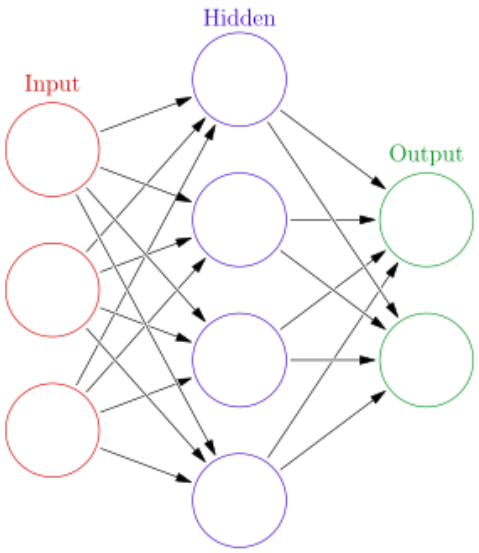
Figura 5 – Ilustração do neurônio artificial. Nela observamos os dados de entrada (*inputs*) e pesos (*weights*). Os valores de entrada e pesos são somados e depois seu valor é analisado por uma função de ativação (*activation function*) que avalia se o valor obtido é acima de um valor de limiar e se positivo envia um sinal (*output*) a outro neurônio.



Fonte: McCullum (2021)

Esses neurônios são agregados em camadas e daí surge o termo *deep learning* ou aprendizado profundo que são rede neurais artificias com múltiplas camadas. Esses algoritmos aprendem através da relação de *inputs* e *outputs* (dados de saída), como cada ligação entre neurônios possui um peso, o treino de redes neurais procura ajustar esse valor com base na diferença entre o valor gerado pela rede neural e o valor real. Assim, os pesos vão se ajustando com base nos dados recebidos e aprendendo padrões a partir do que se espera que seja previsto.

Figura 6 – Camadas de neurônios.



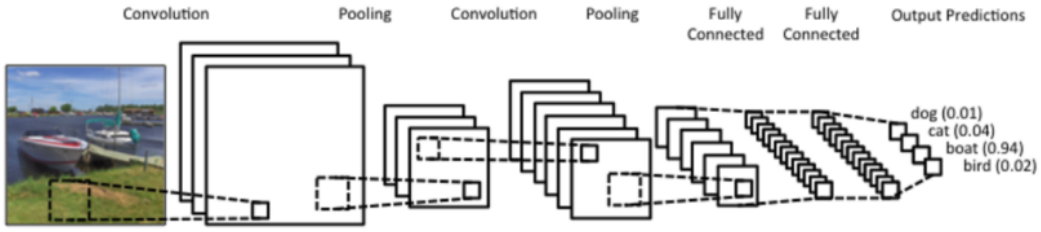
Fonte: McCullum (2021)

Deep learning se mostrou ser muito efetivo ao lidar com dados que são grandes, complexos e não estruturados como textos, imagens e áudio. Seu uso se tornou possível por causa da evolução dos recursos computacionais. Atualmente, ele é utilizado para resolução de tarefas como detecção de objetos ((REDMON et al., 2016),(REN et al., 2016)), reconhecimento facial (ZULFIQAR et al., 2019) e tradução de textos (ZHU et al., 2020).

4.3.1 Redes neurais convolucionais (CNNs)

As redes neurais convolucionais (CNNs) são arquiteturas de deep learning capazes de aplicar filtros (ou também operações de convolução) em larga escala. Elas conseguem aprender representações (*features*) dos dados importantes para a resolução de um problema. Elas consistem na junção de *convolutional layers* com *fully connected layers*. As primeiras extrai as features dos dados de entrada enquanto a última aprende quais são as representações importantes para a resolução do problema. A figura 7 ilustra visualmente a arquitetura. Em visão computacional, essas redes tem sido a base de resolução de muito problemas por conseguirem lidar com uma grande quantidade de dados e aprender a resolver problemas complexos.

Figura 7 – Ilustração de uma CNN.

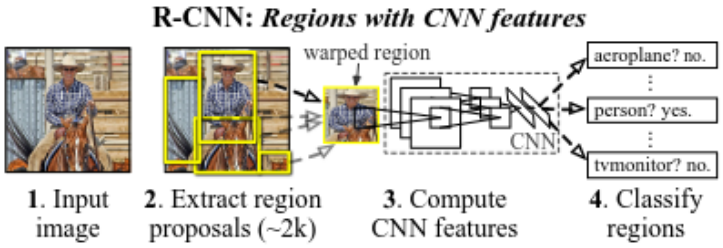


Fonte: Facure (2017)

Além de classificação de imagens, as CNNs também são utilizadas para detecção de objetos. A seguir são listadas outras formas como elas são utilizadas nesse tipo de problema:

- *R-CNN* e *Faster R-CNN*: Em detecção de objetos, as CNNs são utilizadas principalmente para a extração de features das imagens. O algortimo de R-CNN (GIRSHICK et al., 2014) (figura 8) divide a imagem em regiões de interesse (RoI) utilizando *selective search* e aplica as CNNs nelas extraindo representações e depois classificando qual o objeto existente.

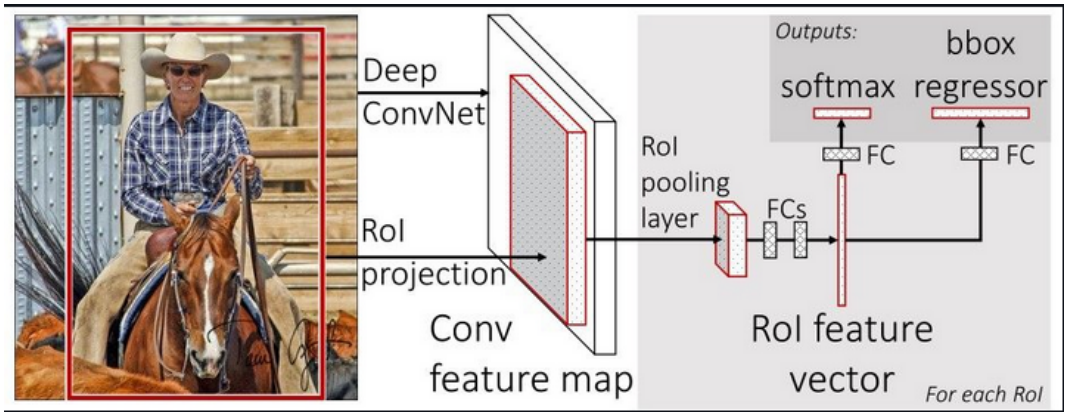
Figura 8 – Ilustração da R-CNN.



Fonte: Girshick et al. (2014)

Uma evolução desse algoritmo é a *Faster R-CNN* (REN et al., 2016) (figura 9) que utiliza uma única CNN para gerar as *bounding boxes* necessárias para detectar objetos. Ela recebe uma imagem e gera representações com a CNN e depois identifica regiões de interesse e as reúne em quadrados, depois elas são transferidas para a *RoI pooling layer* onde são transformadas em vetores de tamanho fixo para serem classificados. O motivo dela ser mais rápida é porque a *RoI pooling layer* permite compartilhar os cálculos realizados em cada RoI em vez de ser um processo separado como no modelo anterior.

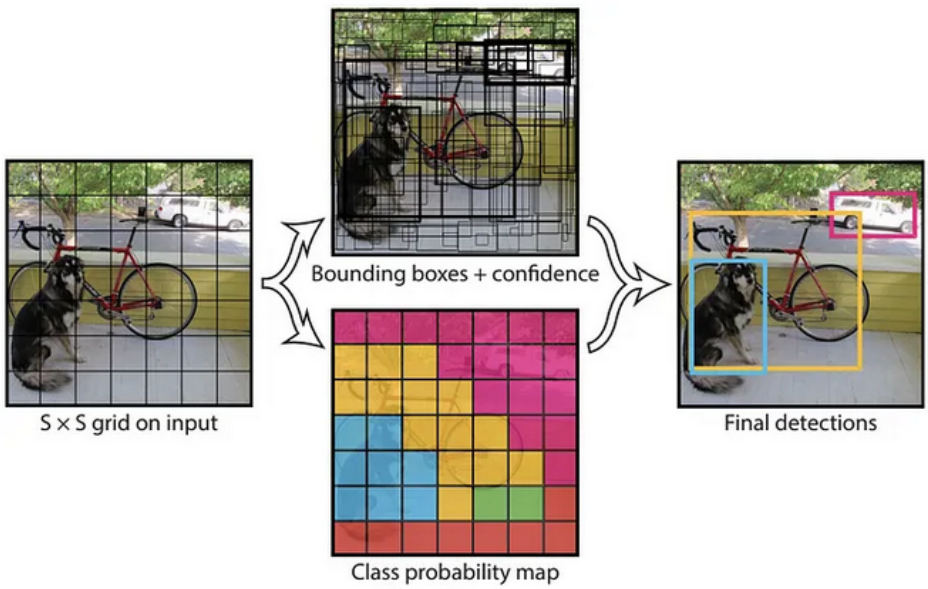
Figura 9 – Ilustração da Faster R-CNN.



Fonte: Gad (2021)

- *YOLO*: A YOLO (*You Only Look Once* (REDMON et al., 2016) ilustrado na figura 10) é um algoritmo de detecção de objetos que se baseia em CNNs puramente para detectar as *bounding boxes*. Ela aplica uma *SxS grid* na imagem e para cada *grid* calcula "m"*bounding boxes*. Em seguida, para cada uma o algoritmo calcula a probabilidade de pertencer a uma classe e os valores da *bounding box*. As que tiverem probabilidade de classe acima de um *threshold* são utilizadas para localizar objetos na imagem.

Figura 10 – Ilustração da YOLO.



Fonte: Gandhi (2018)

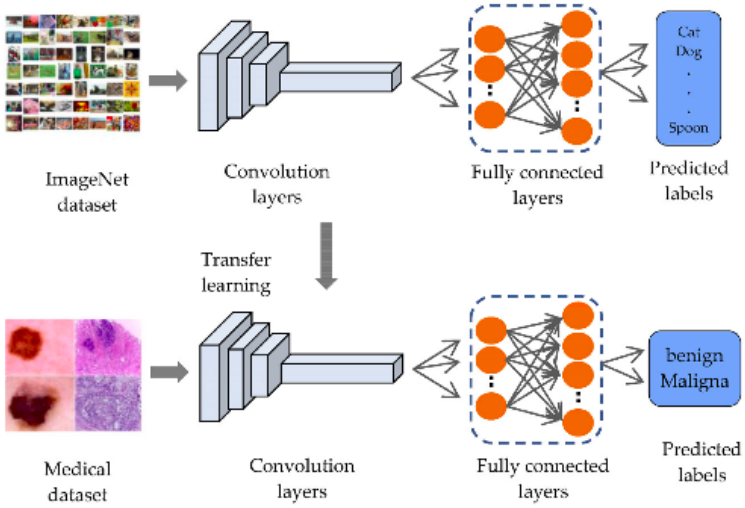
Ela também é rápida assim como os outros algoritmos, contudo, possui dificuldade

para identificar objetos pequenos na imagem, devido a restrições espacias do algoritmo.

4.3.2 *Transfer Learning*

O transfer learning é uma técnica de deep learning muito utilizada para trabalhar com datasets pequenos. Ele consiste em utilizar uma rede pre-treinada e reutilizar sua base convolucional, aplicar os dados de treinamento e treinar uma nova camada de classificação da CNN. A figura 11 ilustra essa abordagem.

Figura 11 – Ilustração de transfer learning para detecção de doenças em imagens médicas.



Fonte: Mukhlif, Al-Khateeb e Mohammed (2023)

As redes pre-treinadas são modelos criados para lidaram com grande quantidade de imagens e resolver problemas mais complexos. Sendo assim, é possível reaproveitar sua base convolucional pois em suas camadas temos aspectos mais generalistas das imagens. As primeiras camadas contém features mais básicas das imagens (bordas, cores, linhas verticais e outros) e nas camadas superiores existem features mais complexas (texturas e localização dos olhos por exemplo) para problemas de visão computacional essas features podem ser reaproveitadas na resolução de outras tarefas.

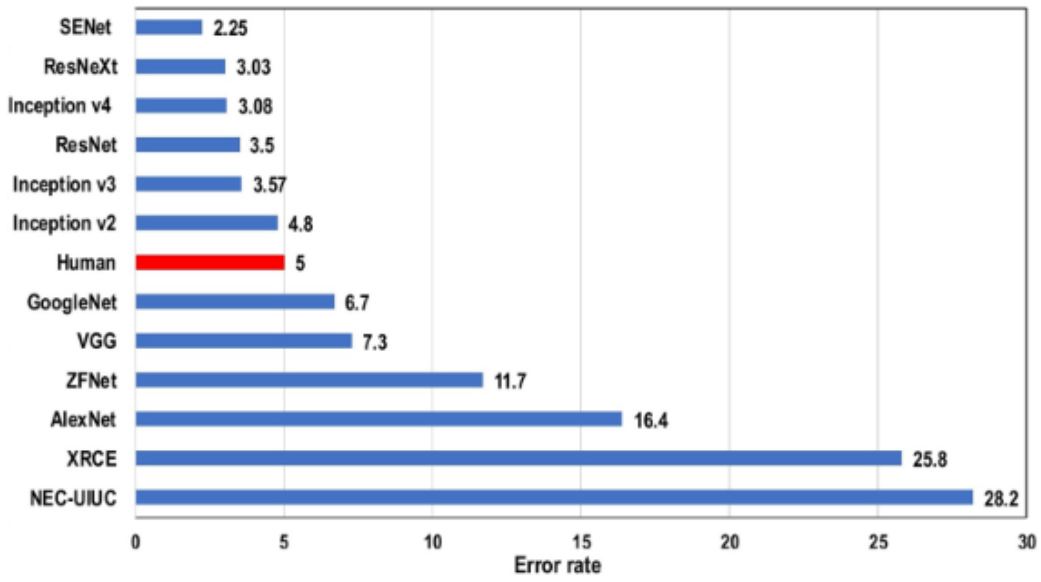
Contudo, a camada de classificação da rede pre-treinada, de forma geral, não pode ser reaproveitada para modelos de CNN. Ela é treinada apenas para resolver um tipo de problema especifico como classificar imagens, por exemplo, o que é diferente de resolver uma tarefa de segmentação de objetos. Então, não é possível reaproveitar seus pesos e por isso deve ocorrer um novo treino para a rede aprender quais das várias features encontradas na rede pre-treinada é relevante para problema a ser resolvido.

4.3.3 Arquiteturas tradicionais

Com o avanço de recursos computacionais e pesquisa, as CNNs se mostraram efetivas na classificação de imagens. Dessa forma, diversos modelos pre-treinados utilizam dessa arquitetura e trazem avanços para a resolução dessa tarefa. O projeto realizador do dataset ImageNet (DENG et al., 2009-) realiza uma competição anual chamada *ImageNet's Large Scale Visual Identity Challenge* (ILSVRC), nela algoritmos são desenvolvidos e testados para reconhecer cenas e objetos. Nessa competição, arquitetura como AlexNet foi a primeira a utilizar deep learning e conseguir obter o menor erro, ela conseguiu tornar padrão tais práticas: *ReLU* como função de ativação nas *hidden layers* e camada de *dropout* onde a partir de um valor numérico ela desativa neurônios ao fazer a classificação, isso força o modelo a aprender através de diferentes combinações deles.

Em seguida, outros modelos vencedores como a VGG (SIMONYAN; ZISSERMAN, 2015), aumentaram a quantidade de camadas e padronizou o uso de *kernel filters* menores. Já a Inception (SZEGEDY et al., 2014), também aumentou as camadas e conseguiu melhorar o resultado através dos *inception modules* que realizam diferentes operações nas features de forma paralela e depois concatenam esses resultados em um único vetor, ela também provou que o uso de camadas menos complexas (*pooling layers*) para realizar a classificação conseguem trazer bons resultados. Depois, a ResNet (HE et al., 2015), um modelo ainda mais profundo, introduziu o conceito de *residual connections* que tornou possível a transferência de conhecimento entre múltiplas camadas além de aplicar a *batch normalization* para estabilizar o treinamento.

Figura 12 – Top 5 *Error Rate* das arquiteturas no dataset ImageNet.



Fonte: Alzubaidi et al. (2021)

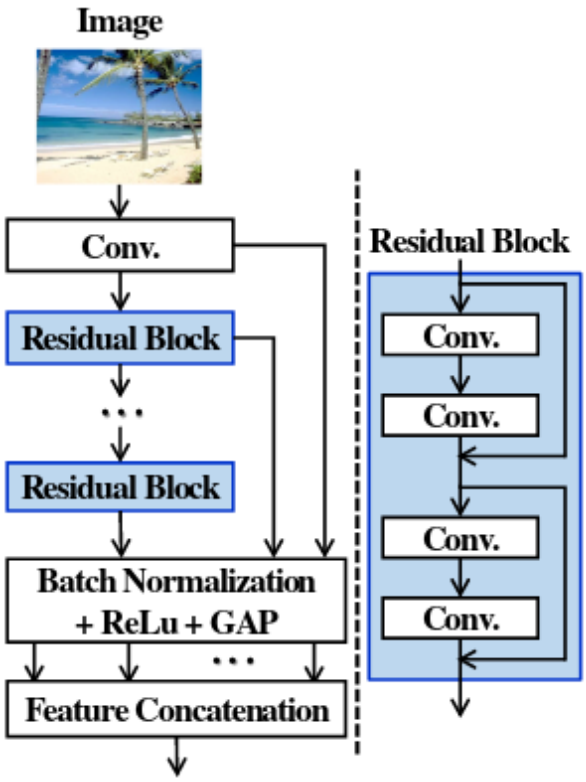
Todas as melhorias realizadas nessas arquiteturas e sua disponibilidade possibilitaram utilizá-las para diferentes tipos de atividades e datasets pequenos. Além disso, também surgem modelos que são mais eficientes em relação ao uso de memória e tamanho como EfficientNet ((TAN; LE, 2020)) e MobileNet ((HOWARD et al., 2017)) sendo possível ser utilizados em dispositivos mobile e aplicações web.

4.4 O problema de classificação de cena

Classificação de cena é uma área da visão computacional que busca categorizar imagens de acordo com seus objetos, conteúdo do ambiente e a maneira como eles estão dispostos na imagem. Dessa forma, as categorias de cenas são mais abrangentes como ”praia”, ”livraria”ou ”floresta”. Diferente da classificação de imagens, onde se está focado em categorizar um único objeto ou um único conceito, classificar uma cena envolve categorizar um contexto ou ambiente representado na imagem. É uma área fundamental para o desenvolvimento de sistemas inteligentes e com uma longa e diversa lista de aplicações como *content based retrival* ((MEHMOOD et al., 2018), (VOGEL; SCHIELE, 2007)), *smart video surveilance* ((SREENU; DURAI, 2019), (KARBALAIE; ABTAHI; SJÖSTRÖM, 2022)), *smart content moderation* ((AKYON; TEMIZEL, 2022),(DEMARTY et al., 2015)), *autonomous driving* ((LORENTE; RIERA; RANA, 2021),(FUJIYOSHI; HIRAKAWA; YAMASHITA, 2019),(SIKIRIC et al., 2020)) e *robot navigation* ((HOU et al., 2019),(ZHANG; YU; HE, 2018)).

O principal objetivo em um problema de classificação de cena é aprender a representação de cena, ou seja, saber quais são as features relevantes para realizar uma categorização. Por isso, muito dos trabalhos nessa área se baseiam em Convolutional Neural Networks (CNNs) justamente por elas conseguirem extrair e aprender features importantes a cada classe. Com o advento dos grandes datasets ImageNet ((DENG et al., 2009-)) e Places ((ZHOU et al., 2018)), primeiramente, se pre-treinava um modelo neles e depois se realizava o transfer learning para um dataset menor e apenas com cenas. Através do avanço da pesquisa, outras técnicas passam a ser utilizadas com CNNs para melhorar a acurácia, o modelo FTOTLM de (LIU; TIAN; XU, 2019) (figura 13) usa dos *residual blocks* introduzidos pela ResNet (ZENG et al., 2021) para extrair múltiplas features de uma imagem e concatena-las em um único vetor para posterior classificação.

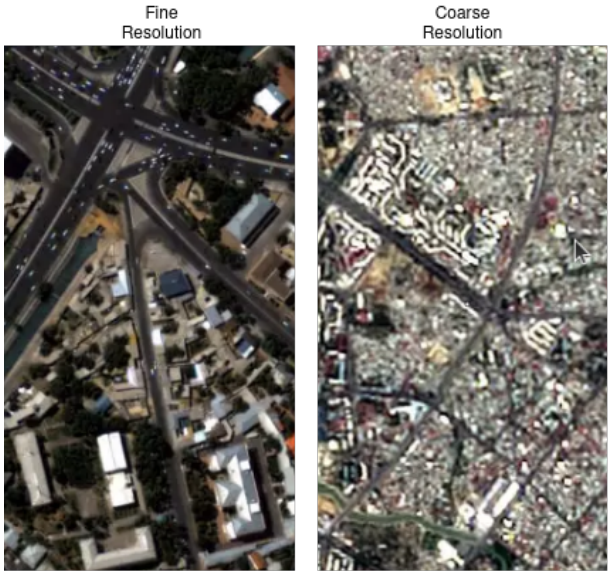
Figura 13 – Ilustração da arquitetura FTOTLM.



Fonte: Zeng et al. (2021)

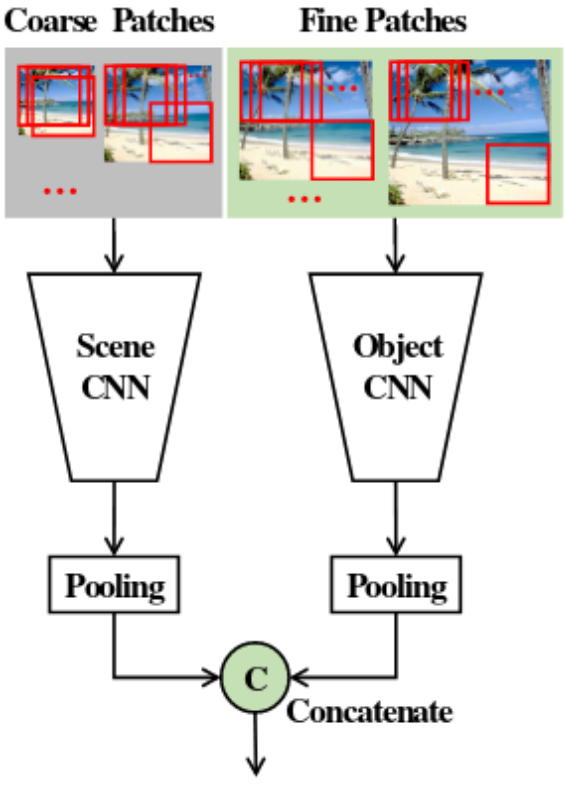
Uma outra forma de aplicação é vista na arquitetura de (HERRANZ; JIANG; LI, 2016) (figura 15). Ela preprocessa a imagem da cena em múltiplas escalas e com diferentes resoluções, depois as aplica em duas CNNs. As imagens em *coarse resolution* tem features extraídas pela *Scene CNN*, um modelo focado em extrair informações de contexto da imagem, já a as imagens em *fine resolution* passam pela *Object CNN* a qual extrai informações sobre os objetos na imagem, por fim, ambas essas informações são concatenadas em um vetor para serem classificadas.

Figura 14 – Ilustrações de coarse resolution e fine resolution. A esquerda (fine resoluiton) mostram imagens com alta nitidez e riqueza de detalhas, isso porque cada pixel abrange uma área pequena (entorno de 1m a 10m). Já a imagem da direita (coarse resolution), cada pixel engloba uma área grande (maiores que 30m).



Fonte: Própria

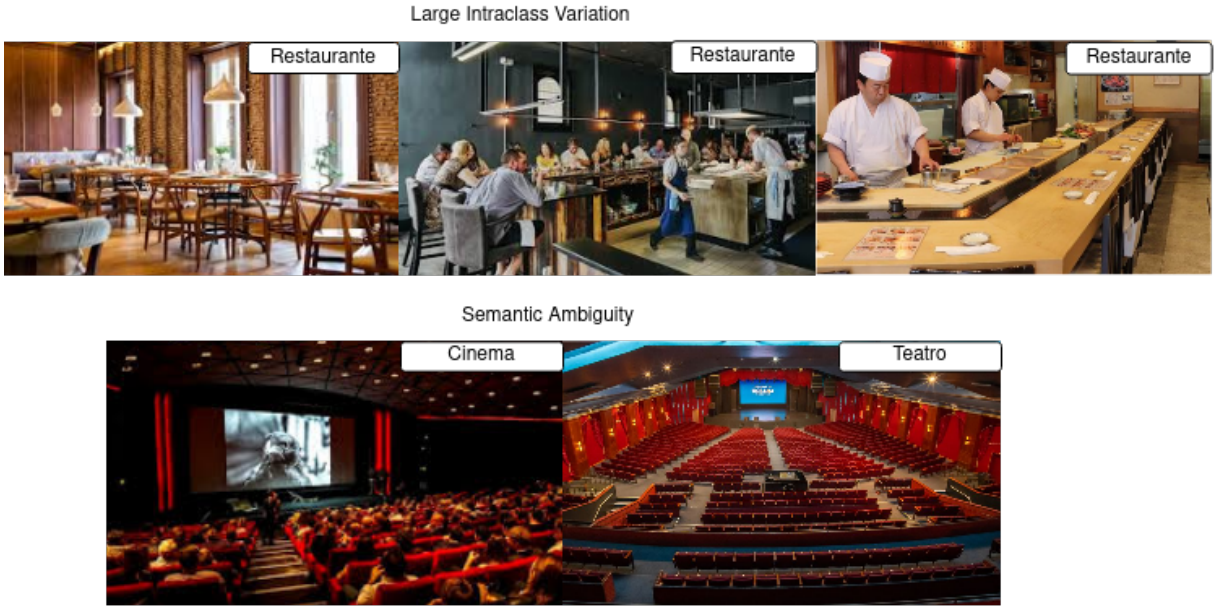
Figura 15 – Arquitetura de Herranz, Jiang e Li (2016)



Fonte: Herranz, Jiang e Li (2016)

Contudo, apesar dos avanços ainda existem dificuldades na resolução de classificação de cenas. Um deles é a grande variação entre as classes (*large intra class variation*), isso significa que uma mesma classe pode conter diferentes objetos,backgrounds ou atividades humanas, além disso, diferenças em qualidade de imagem como pouco/muita luz, alta/baixa resolução pode afetar a criação de uma representação para uma classe. Outro problema, é a ambiguidade semântica (*semantic ambiguity*), imagens de diferentes classes podem conter objetos iguais com relações semelhantes entre si, ou seja, visualmente classes distintas são parecidas, dessa forma, torna-se difícil para o modelo diferenciar tais cenas. A figura 16 demonstra os problemas citados. Por causa disso, as arquiteturas comentadas apresentam dificuldades em classificar imagens de natureza, principalmente as de sensoriamento remoto.

Figura 16 – Ilustrações de large intraclass variation e semantic ambiguity. No topo (*large intraclass variation*), observamos que restaurantes apresentam uma grande diversidade de layouts e objetos podendo variar de acordo com a cozinha feita. No fundo (*semantic ambiguity*), verifica-se que cinema e teatro apresentam ambientes bem similar com mesmas cores, objetos e layout na cena.



Fonte: Própria

Imagens de sensoriamento remoto são imagens da superfície da Terra obtidas a uma distância, elas são adquiridas por satélite, drone ou avião. Através delas podemos entender melhor o meio ambiente que vivemos e assim é possível monitorar o crescimento de cidades.((BEGNINI, 2023),(YEH et al., 2020),(YU et al., 2022), (SHAFIQUE et al., 2022),(MINETTO et al., 2021)), clima ((Prabhat et al., 2021),(Jacques-Dumas et al., 2022),(MORAUX et al., 2019)) e uso de terra ((CASTELLUCCIO et al., 2015), (Carranza-García; García-Gutiérrez; RIQUELME, 2019),(FENG et al., 2020),(SINGH et al., 2022)).

Dada a importância, datasets e modelos para classificar cenas tem sido mais desenvolvidos recentemente. Inicialmente, optou-se por trabalhar com CNNs da maneira tradicional, inclusive utilizando modelos pre-treinados com ImageNet, todavia, além das dificuldades citadas anteriormente, é visto que imagens de sensoriamento remoto se diferenciam muito das existentes na ImageNet. As primeiras, contêm em uma imagem mais de um objeto, com eles distantes entre si e também apresentam classes muito desbalanceadas. Outro desafio também é obter dados rotulados para o treinamento dos modelos. Geralmente, as imagens são obtidas sem rótulos sendo necessário um trabalho manual de rotulação para depois aplicar um modelo, isso muitas vezes é custoso e lento.

Figura 17 – Ilustrações de large intraclass variation e semantic ambiguity em imagens de sensoriamento remoto (*remote sensing*). No topo (large intraclass variation), observamos que terras agrícolas apresentam uma grande diversidade de formatos e áreas cobertas. No fundo (semantic ambiguity), verifica-se que playground e estádio podem ter elementos similares, geralmente relacionados ao esporte, como campo de futebol ou pistas de atletismo.

Fonte: Própria

Dessa forma, técnicas que consigam obter bons resultados para classificação utili-

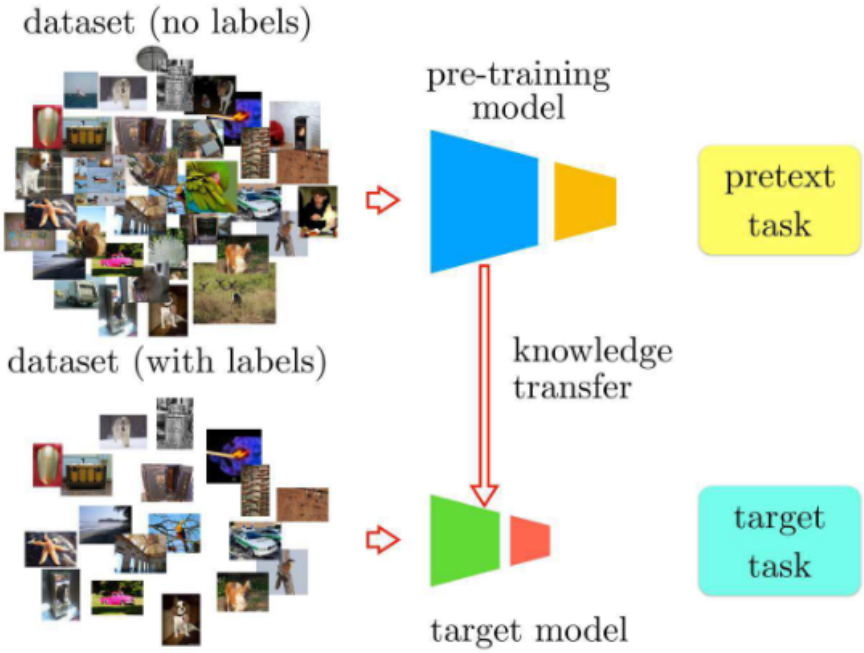
zando poucos dados rotulados tem sido aplicadas. A principal delas, é conhecida como self-supervised learning. Ela é conhecida por ser um novo paradigma de treinamento de modelos. Primeiro, um modelo aprende a criar representações das classes do dataset através da *pretext task*, depois, a partir das representações é treinado um novo modelo para a *target task*. Em conjunto com isso, para melhorar a acurácia dos modelos de classificação de cena tem se trabalhado com mais de uma modalidade de dados, sendo então, modelos multimodais. A vantagem de trabalhar com mais de um tipo de dado é justamente oferecer mais contexto de um problema para o modelo classificar corretamente.

4.5 Novas técnicas

4.5.1 *Self-supervised learning*

Self-supervised learning (SSL) é um paradigma em machine learning no qual o modelo se utiliza dos dados sem labels para aprender representações deles e depois aplicá-las para resolução de tarefas de classificação, regressão ou clustering. Isso é feito através da geração automática de pseudolabels (labels que refletem atributos dos dados), como rotação ou filtros aplicados a imagens, ou a reordenação de palavras em textos. Ao aplicar um treinamento supervisionado para prever essas pseudolabels (pretext tasks), o modelo adquire características dos dados e constrói representações significativas para utilização em outras tarefas (target tasks). A figura 18 ilustra esse método.

Figura 18 – Ilustração do método self-supervised learning.



Fonte: Noroozi et al. (2018)

A vantagem de se adotar a abordagem de treinamento utilizando SSL é o fato dela aprender representações relevantes dos dados mesmo sem labels. Apesar da grande disponibilidade de dados, o processo de realizar anotações e definir labels ainda é lento e custoso. Por isso, formas alternativas de treinamento, como a SSL, tem sido estudadas e testadas para obter melhores modelos para tarefas de visão computacional e que não dependem de uma grande quantidade de dados anotados.

Quando se utiliza SSL é importante escolher a pretask task correta, caso contrário, modelo não consegue aprender tão bem as representações a partir das pseudolabels. Dependendo da target task muito das transformações feitas nos dados para gerar as labels pode causar perdas de informações cruciais para diferenciar classes. A exemplo, ter pretasks relacionadas a coloração de imagens pode ser útil para classificar objetos muito distintos como ”floresta”e ”terras agrícolas”, porém, torna difícil uma classificação mais refinada como definir tipos de floresta ou terras agrícolas conforme ilustra a figura 19 abaixo.

Figura 19 – Ilustração da escolha da pretask task correta. Pelas imagens vemos que floresta e terra agrícola têm características distintas, contudo, quando observamos os tipos de ambientes as cores se tornam elementos importantes para distinguir as classes.

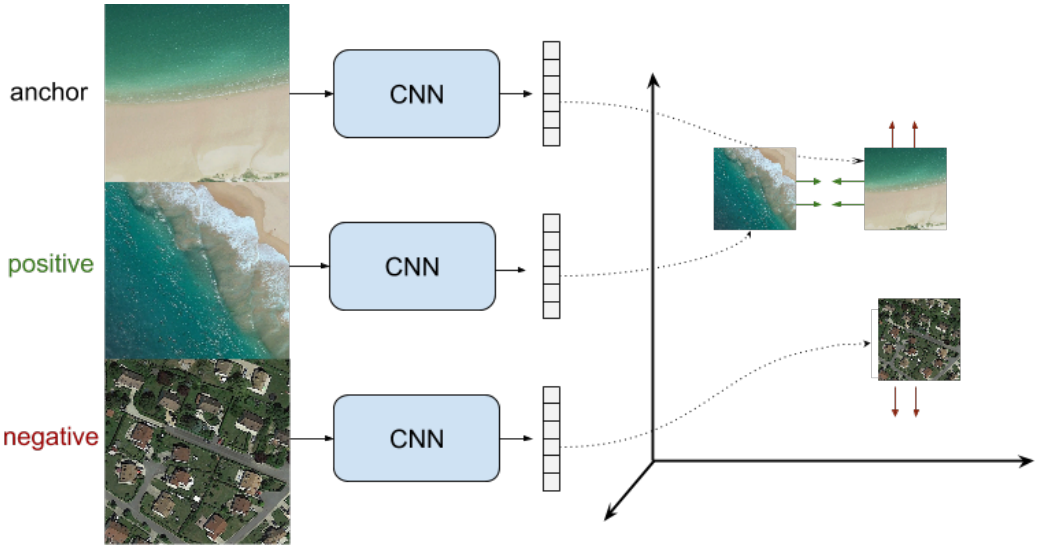
Fonte: Própria

Portanto, para criação de representações melhores com imagens de sensoriamento remoto, atualmente, tem-se empregado o método de SSL com *constrative learning* ((STOJNIĆ; RISOJEVIĆ, 2021),(BERG; PHAM; COURTY, 2022)). Nessa abordagem, o modelo re-

cebe pares de dados, os quais podem ser positivos, se forem da mesma classe ou negativos, se forem de classes diferentes. Assim, ele aproxima as representações dos pares positivos e as afasta dos pares negativos. Para fazer isso, uma das formas mais simples e utilizadas é a *triplet loss* (BALESTRIERO et al., 2023). Ela é conhecida pela equação 1, dado um item de âncora e seu par positivo e negativo, ela procura deixar a representação do item de âncora mais próxima do seu par positivo e o mais distante (acima de uma margem "m") do seu par negativo.

$$L(r_a, r_p, r_n) = max(0, d(r_a, r_p) - d(r_a, r_n) + m) \tag{1}$$

Figura 20 – *Triplet loss* para imagens de sensoriamento remoto.

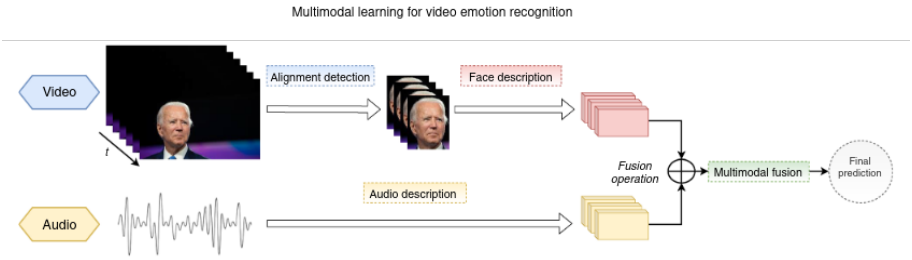


Fonte: Própria

4.5.2 *Multimodal learning*

Multimodal learning se refere a construção de modelos que combinam informações vindas de múltiplas modalidades de dados, isto é, dados de diferentes fontes. Essas fontes podem ser imagens, texto, áudio, dados estruturados e até mesmo dados de sensores. O objetivo disso é melhorar a performance dos modelos para tarefas como descrição de imagens/vídeos, classificação de cenas/emoções e reconhecimento de fala. Ao reunir diferentes tipos de dados é possível obter um entendimento mais profundo de fenômenos complexos e então melhorar resultados em várias aplicações.

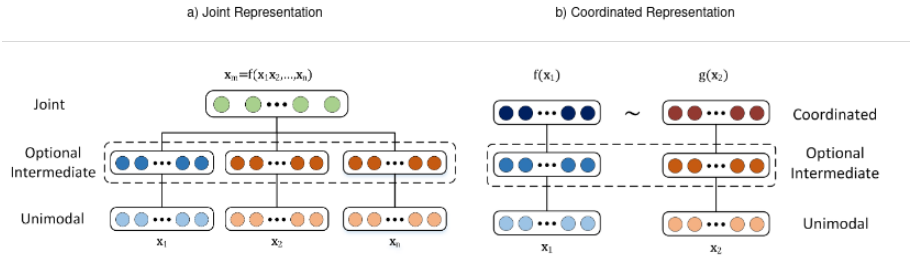
Figura 21 – Exemplo de arquitetura de *multimodal learning* para *video emotion recogniton*.



Fonte: Baltrušaitis, Ahuja e Morency (2017)

Ao trabalhar com modelos multimodais um dos principais desafios é criar representações que combinem informações de fontes distintas de dados. Duas formas de fazer isso é através *joint representation* e *coordinated representation* as quais são descritas abaixo (e ilustradas na imagem 22):

Figura 22 – Em a) Ilustração de *joint representation learning* e b) Ilustração de *coordinated representation learning*.



Fonte: Baltrušaitis, Ahuja e Morency (2017)

- *Join representation*: Ela combina as representações de cada fonte de dados, concatenando-as em um vetor por exemplo. Outras operações podem ser feitas também como adição ou subtração.
- *Coordinate representation*: cria representações distintas para cada fonte de dados, contudo, elas são coordenadas através de restrições. A exemplo, uma restrição é fazer com que a representação da palavra ”cachorro”e imagem de um ”cachorro”sejam próximas entre si, enquanto, a palavra ”cachorro”e imagem de um ”gato”são distantes entre si.

Qual dessas formas escolher depende dos dados disponíveis no momento de inferência. Quando se têm todas as modalidades disponíveis, se prefere usar joint representation learning, por isso ela tem sido muito utilizada para tarefas como *audio-visual scene recognition*, *audio-visual speech recognition*, *multimodal gesture recognition* e *multimodal event detection*. Já a coordinate representation learning é utilizada quando se tem apenas

uma das modalidades disponivel na inferência como ocorre em *image captioning*, *video description* e *cross-modal retrieval*.

Recentemente, no contexto de sensoriamento remoto também tem se utilizado abordagens multimodais para tarefas como classificação de imagens (LI et al., 2022) e land use ((SRIVASTAVA; Vargas-Muñoz; TUIA, 2019)). Nesse domínio é conhecido múltiplos tipos de dados como *multi-spectral*, *hyperspectral* e SAR (*synthetic-aperture radar*). Contudo, por essas abordagens trabalharem apenas com imagens, quando aplicadas no mundo real elas encontram dificuldades devido a mudança temporal do espaço, má qualidade de imagem e diferentes fontes de dados (como satélites e drones). Para contornar isso, devido ao uso massivo de *smartphones*, *wearables* e plataformas de compartilhamento de áudio é possível obter dados de áudio geo-localizados e dessa forma obter datasets que combinem dados de imagens remotas e áudio de diversos locais do planeta. Os trabalhos de (HU et al., 2020) e (HEIDLER et al., 2021) reúnem esses dados e mostram o potencial de combinar essas informações para tarefas como classificação audiovisual de cenas e *cross-modal retrieval*.

4.6 Métricas de avaliação

As métricas de avaliação são cálculos realizados a partir dos valores previstos pelo modelo e os valores reais. A partir delas é possível avaliar quão bem o modelo está aprendendo. Para problemas de classificação, as métricas mais importantes são as listadas abaixo:

- Matriz de confusão ou *Confusion Matrix*: É uma matriz que reúne a quantidade de dados previsto pelo modelo e os valores reais. Ela é utilizada para verificar a frequência de acertos e erros. (GÉRON, 2019)

Tabela 1 – Matriz de Confusão

	Previsto	
Real	Positivo	Negativo
Positivo	TP	FN
Negativo	FP	TN

Na Tabela 1, é representada uma matriz de confusão, onde que:

  - Verdadeiro Positivo (*"True Positive"* - TP): é a quantidade de dados que o modelo previu a condição como positiva e era realmente positiva;
  - Falso Positivo (*"False Positive"* - FP): é a quantidade de dados que o modelo previu a condição como positiva e, na verdade, era negativa;

  - Verdadeiro Negativo (*"True Negative"* - TN): é a quantidade de dados que o modelo previu a condição como negativa e realmente era negativa;
  - Falso Negativo (*"False Negative"* - FN): é a quantidade de dados que o modelo previu a condição como negativa e, na verdade, era positiva.

- Acurácia ou *Accuracy*: É média de acertos do modelo. É a relação entre os valores corretos previstos e número total de previsões.(GÉRON, 2019)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

- Precisão ou *Precision*: Ela mede em relação aos valores previstos quantos deles estão corretamente classificados pelo modelo. (GÉRON, 2019)

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

- *Recall* ou *Sensitivity*: Ela mede em relação aos valores reais quantos deles estão corretamente classificados pelo modelo. (GÉRON, 2019)

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

- *F1-Score*: Ela é a média harmônica entre recall e precision. Seu valor máximo é quando recall e precision tem valores iguais. (GÉRON, 2019)

$$\text{F1-score} = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

5 METODOLOGIA

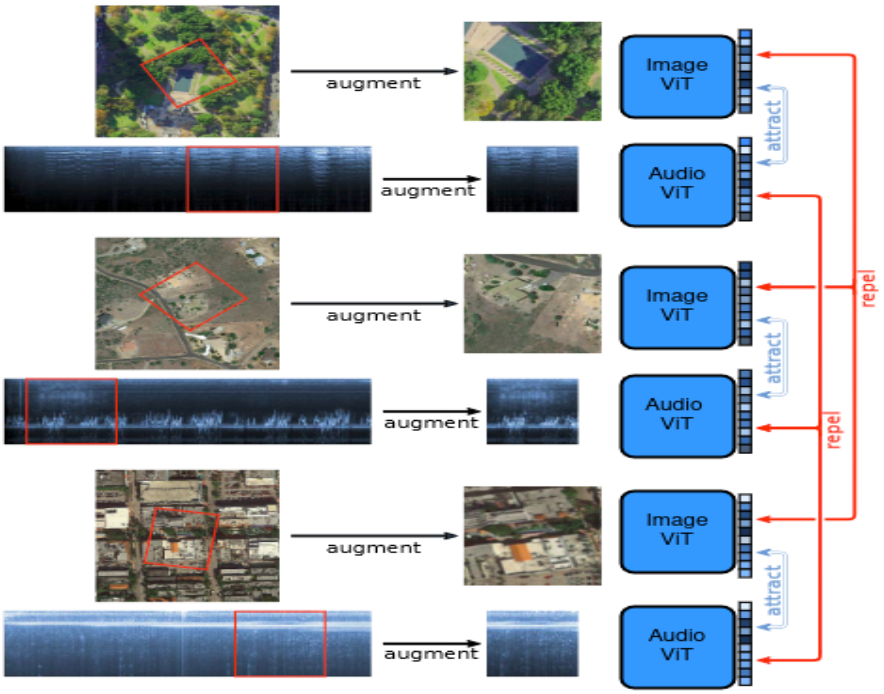
O problema de classificação de cena vem ganhando notoriedade dado a sua importância e variedade de aplicações. Contudo, obter dados rotulados é um processo lento e custoso, por isso, a pesquisa se dirigiu as técnicas de self-supervised learning.

Além disso, por causa de uma grande quantidade de dados de imagem e som disponíveis atualmente. Pesquisas como (HU et al., 2020) e (HEIDLER et al., 2021) introduziram datasets com essas modalidades para obter melhores performances de modelos de deep learning para resolução desse problema.

O trabalho de (HEIDLER et al., 2021) proprõem uma arquitetura que utiliza do método do *constrative self-supervised learning* para aprender representações de dados audiovisuais de sensoriamento remoto. Nela, tanto imagens quanto áudios são codificados por CNNs e então comparados utilizando a *batch triplet loss*.

Porém, recentemente, os *vision transformers* se mostraram promissores por criarem representações ricas em contexto e conseguirem capturar relações de longa distância em tempo e espaço que são importantes para capturar informações de movimento. Logo, o presente trabalho pretende se utilizar deles para melhorar os *embeddings* (representações) de imagens gerados pelo modelo proposto por (HEIDLER et al., 2021) e aplicar esse mesmo método para realizar a classificação de cenas com dados audiovisuais de sensoriamento remoto utilizando os datasets *Sounding Earth* e *ADVANCE*.

Figura 23 – Metodologia proposta.



Fonte: Própria

5.1 Datasets

Os datasets utilizados para realizar a classificação de cena são os listados abaixo:

- *Sounding Earth dataset*: Esse é o dataset pra ser utilizado para pretask. Ele consiste em 50.545 pares de imagem e áudio. Os áudios geo-localizados foram obtidos através dos dados disponíveis pelo projeto Radio Aproee ::: Maps que coletou cerca de 50.000 áudios de várias localidades do mundo. A partir das informações de latitude e longitude dos sons, uma imagem de 1024 x 1024 pixels referente ao local é extraída em alta-resolução pelo Google Earth.
- *ADVANCE dataset*: Esse já é um dataset com labels de 13 classes: *airport, sports land, beach, bridge, farmland, forest, grassland, harbor, lake, orchard, residential area, shrub land, and train station.* Nesse caso, ele é utilizado como *benchmark* para o problema de classificação audiovisual de cenas aéreas.

5.2 Preprocessamento dos dados

- Dados de áudio: Para processar os dados de áudio, primeiro, eles foram transformados em uma imagem representando o espectro em frequência (espectrograma) do sinal pela *short term Fourier transform* (STFT). Em seguida, o valor ao quadrado do valor absoluto dos coeficientes da transformada foi mapeada para a escala "mel"(*mel scale*) utilizando 128 filtros de banda. Assim, é obtido uma imagem 128 x T onde T é o valor do tamanho do áudio. Por fim, para padronizar o tamanho dos espectros, em cada um é retirado 128 frames consecutivos, resultando em um espectrograma de tamanho 128 x 128 pixels. Contudo, para se aplicar os embeddings com vision transformers foi necessário realizar a conversão da imagem para RBG e também aumentar seu tamanho para 224 x 224.
- Dados de imagem: Nas imagens por elas serem de tamanho grande e de longa distância, primeiro, o centro da imagem é cortado, gerando uma imagem de tamanho entre 192 e 394 pixels que depois é reduzida para um tamanho de 224 x 224 pixels. Ao final, outros tratamentos são aplicados como rotação, blur, hue, saturação e ajuste de cores.

5.3 *Embeddings Networks*

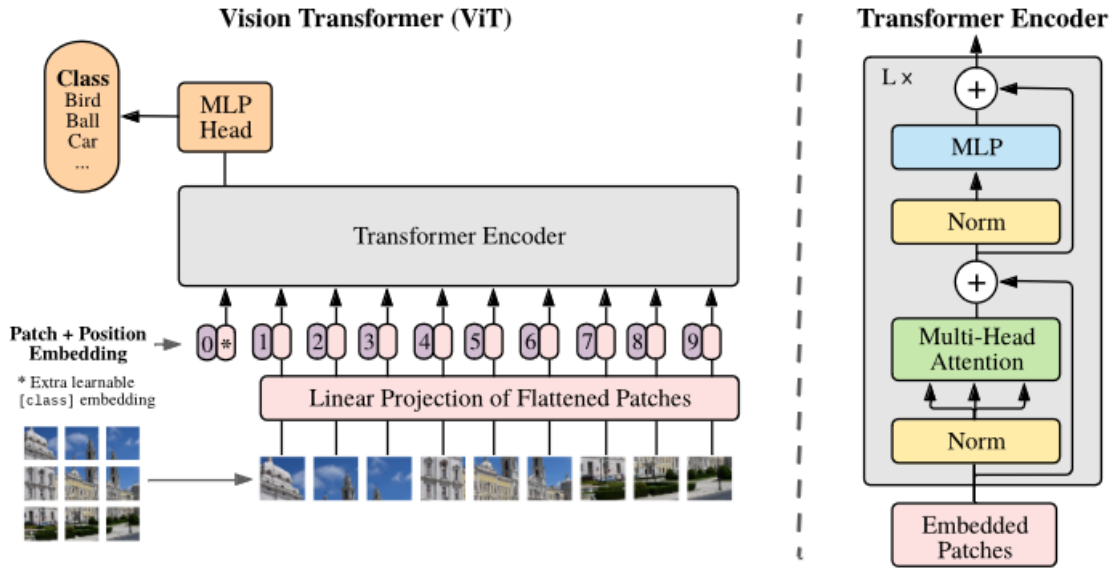
Tanto imagens quanto as *waveforms* de sons compartilham atributos e propriedades diferentes. Por isso, para cada tipo de dado se faz necessário codificar essas informações em embeddings, isso é, projetar a grande quantidade de informações em vetores menores.

Dessa forma, a arquitetura proposta possui duas embeddings networks: *audio subnet* e *image subnet*. Elas procuram aprender quais são as correlações entre os dados de diferentes modalidades.

- *Audio subnet*: possui dados de entrada, espectrogramas log-mel, com tamanho 3 x 224 x 224. Possui um vision transformer (ViT) encoder e ao final uma CNN (em uma única dimensão) e fully connected layer são aplicadas para obter um único vetor representando um áudio.
- *Image subnet*: possui dados de entrada com tamanho 3 x 224 x 224, também possui o vision transformer (ViT) encoder e ao final uma CNN (em uma única dimensão) e fully connected layer são aplicadas para obter um único vetor representando uma imagem.

É valido destacar aqui o uso do vision transformer como encoder das imagens. Ele é conhecido por um tipo de rede neural que aprende quais são as propriedades mais importantes dos dados através do seu mecanismo de *self-attention*. Conforme mostra a figura 24 a imagem é dividida em *patches*, depois, cada uma é projetada para uma dimensão menor e adicionada a um número referente a sua posição na imagem (*position embedding*), por último, elas são reunidas e codificadas pelo transformer encoder, nele é calculado qual o grau de relacionamento de cada patch entre si e mede-se o quão isso é importante para realizar a classificação na camada final.

Figura 24 – Ilustração do Vision Transformer (ViT).



Fonte: Dosovitskiy et al. (2021)

Com uma grande quantidade de dados os vision transformers aprendem características específicas de imagens sem depender de redes específicas como CNNs e resultar em arquiteturas com menos complexas.

O resultado final de cada um dos modelos são vetores que definem a representação

de cada uma das modalidades. O método de treinamento adotado é com self-supervised learning, então, em seguida, eles são avaliados pela batch triplet loss onde o modelo aprende quais representações devem ser próximas entre si e quais devem ser distantes.

5.4 *Batch triplet loss*

Técnicas mais tradicionais para calcular a triplet loss utilizam de dois a três samples de dados, contudo, para ser possível ter mais feedback ao modelo, tem-se adotado utilizar todos os possíveis pares de embeddings em um único batch de treino. Assim, se tem a batch triplet loss definida abaixo.

$$D = \begin{pmatrix} \|a_1 - v_1\|_2 & \cdots & \|a_1 - v_n\|_2 \\ \vdots & \ddots & \vdots \\ \|a_n - v_1\|_2 & \cdots & \|a_n - v_n\|_2 \end{pmatrix} \tag{6}$$

$$\begin{aligned} L(D) = &\sum_i \sum_{j \neq i} \max(0, D_{ii} - D_{ij} + 1) \\ &+ \sum_i \sum_{i \neq j} \max(0, D_{ii} - D_{ij} + 1) \end{aligned} \tag{7}$$

Conforme 6 e 7, ela é definida por uma matriz onde cada valor representa a distância entre um sample de imagem e áudio. O objetivo dela então é minimizar os valores da diagonal principal e deixar os demais valores acima de uma determinada margem.

5.5 Aplicação no ADVANCE dataset

Depois do modelo obtido pelas embeddings networks é possível obter representações ricas em detalhes de imagens e espectrogramas de áudio. Sendo assim, é possível utilizar os embeddings gerados para servir como extração de features do dataset ADVANCE.

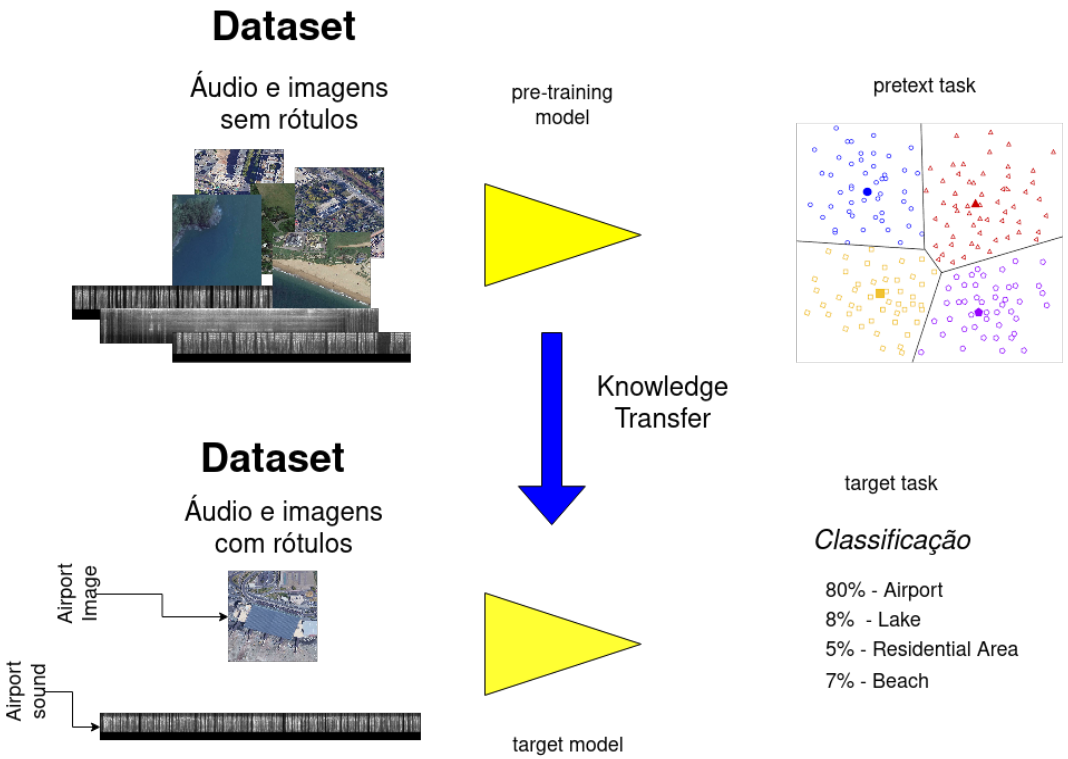
Similar ao Sounding Earth, o ADVANCE também classifica dados de imagens de sensoriamento remoto e áudio geo-localizados, contudo, ele possui labels. Logo, uma regressão logística é treinada nesses dados a partir das representações geradas pelo modelo treinado com a batch triplet loss.

5.6 Ilustração do método para realizar classificação de cenas

Com base no citado anteriormente, a figura 25 ilustra o processo a ser realizado para aplicar o método de self-supervised learning para realizar classificação de cenas aéreas. As representações são aprendidas utilizando um dataset sem labels utilizando a abordagem de constrative learning (batch triplet loss), em seguida, elas são aplicadas em um

dataset já rotulado e um novo modelo é treinado de maneira supervisionada para realizar a classificação de cena.

Figura 25 – Ilustração da metodologia para classificação de cena aéreas



Fonte: Própria

6 RESULTADOS E DISCUSSÕES

Na aplicação do projeto foi utilizado o modelo *timm/vit_base_patch16_224.orig_in21k* e *timm/vit_base_patch32_224.orig_in21k* disponível no model hub da plataforma HuggingFace. Esses modelos são as versões do ViT da imagem 24 pre-treinadas no dataset ImageNet21k. Cada um contém 12 blocos de *attention heads* e tamanho das camadas de *hidden size* de 768 e a diferença entre cada um é a quantidade de patches da imagem (16 e 32).

Para o aprendizado dos embeddings no SoundingEarth foi feito o fine tuning do modelo considerando 15 épocas e batch size de 32 utilizando a GPU V100 disponível no google colab. Já para a classificação de cena no dataset ADVANCE foi considerado 4 tipos de embeddings: apenas imagens, apenas áudio, a média de imagem e áudio e concatenação de imagem. Dessa forma, o modelo de regressão logística é aplicado considerando cada tipo em 5 treinos distintos e a média dos resultados é reunido na tabela 3.

A seguir vamos descrever os resultados analisando o tamanho dos modelos, os scores obtidos para as métricas de precision, recall e F1-Score e por último analisando os resultados por classes.

6.1 Tamanho dos modelos

Tabela 2 – Comparação de tamanho dos modelos.

	Total parameters (M)	Total size (MB)
ResNet18	38.37	94.94
ResNet50	154.46	148.51
ViT 16 patches	181.37	687.05
ViT 32 patches	184.65	700.19

Fonte: Própria

A tabela 2 reúne a informação dos tamanhos dos modelos comparados. Os ViT apresentam muito mais parâmetros. Segundo (DOSOVITSKIY et al., 2021) um único ViT contém mais de 80 milhões de parâmetros, portanto, como são necessários 2 para construção dos embeddings, ao final obtemos um número maior.

6.2 Scores para os modelos

Tabela 3 – Resultado de Precision, Recall e F1-Score no ADVANCE Dataset.

model name	mode	F-score	Precision	Recall
ResNet 18 baseline	concat	**89.50**	**89.59**	**89.52**
	image	**86.92**	**87.09**	**87.07**
	sound	37.69	38.36	37.91
ResNet 50 baseline	concat	88.83	88.90	88.85
	image	83.84	83.97	83.88
	sound	39.01	39.13	39.96
ViT 16 patches	concat	86.34	86.54	86.48
	image	82.60	82.96	83.03
	sound	40.43	41.14	41.26
ViT 32 patches	concat	86.03	86.26	86.17
	image	81.18	81.65	81.50
	sound	**42.56**	**44.64**	**42.90**

Fonte: Própria

Considerando apenas os embeddings de imagem, é visto que a ResNet ainda apresenta um desempenho superior na classificação das classes. Isso significa que os modelos

apenas com CNNs conseguem aprender propriedades importantes das imagens mesmo com uma pequena quantidade de dados, é comum observar tal resultado em arquiteturas convolucionais por elas terem características internas que tornam esse aprendizado mais eficiente. Contudo, é interessante verificar que os ViTs mesmo com um desempenho inferior conseguem também ficar na casa dos 80%, ou seja, apesar do dataset da pretext task ser pequeno, essa arquitetura conseguiu obter embeddings representativos e ter um bom resultado ao classificar cenas, logo, é possível que com mais épocas de treino ela consiga ter um desempenho próximo ou superior ao baseline.

Por outro lado, em consideração as representações dos sons, é visto que os embeddings dos ViTs conseguiu obter uma melhor generalização. Dados de espectrogramas não são tão complexos quanto imagens RGB, não possuem objetos ou texturas, são gráficos. Dessa forma, os vision transformers com o mecanismo de *self-attention* apresentam um ganho considerável em construir representações de áudio melhores.

Em todos os modelos, os melhores resultados para a classificação de cenas no ADVANCE dataset foi com ambos os dados de som e imagem sendo utilizados. Apesar do ResNet se mostrar superior ao ViT, isso se dá, pois os seus embeddings para as imagens nesse caso também são melhor seletores de classes. Todavia, ao analisar o resultado no ViT, é visto que o ganho obtido na representação de som também melhora o resultado final: tanto o ViT com 16 e 32 patches ficaram próximos ao modelos com ResNet18 e ResNet50, apesar de poucas épocas de treino.

6.3 Resultados por classes

A seguir é ilustrado em tabelas e gráfico a quantidade de acerto médio por classe e tipo de embedding. Os valores estão entre 0 e 1, quanto mais próximo a média é 1, maior é a quantidade de acerto para aquela classe.

6.3.1 Média de acerto por classe

Tabela 4 – Estatísticas por classes para ViT 16 patches para image e sound.

mode	class name	count	mean	std	min	25%	50%	75%	max
image	airport	185.00	0.94	0.25	0.00	1.00	1.00	1.00	1.00
	beach	215.00	0.69	0.46	0.00	0.00	1.00	1.00	1.00
	bridge	280.00	0.71	0.45	0.00	0.00	1.00	1.00	1.00
	farmland	430.00	0.80	0.40	0.00	1.00	1.00	1.00	1.00
	forest	865.00	0.90	0.31	0.00	1.00	1.00	1.00	1.00
	grassland	150.00	0.39	0.49	0.00	0.00	0.00	1.00	1.00
	harbour	510.00	0.81	0.39	0.00	1.00	1.00	1.00	1.00
	lake	355.00	0.81	0.39	0.00	1.00	1.00	1.00	1.00
	orchard	205.00	0.78	0.42	0.00	1.00	1.00	1.00	1.00
	residential	1055.00	0.98	0.15	0.00	1.00	1.00	1.00	1.00
	sparse shrub land	405.00	0.84	0.37	0.00	1.00	1.00	1.00	1.00
	sports land	160.00	0.61	0.49	0.00	0.00	1.00	1.00	1.00
	train station	260.00	0.73	0.45	0.00	0.00	1.00	1.00	1.00
sound	airport	185.00	0.55	0.50	0.00	0.00	1.00	1.00	1.00
	beach	215.00	0.33	0.47	0.00	0.00	0.00	1.00	1.00
	bridge	280.00	0.50	0.50	0.00	0.00	0.50	1.00	1.00
	farmland	430.00	0.26	0.44	0.00	0.00	0.00	1.00	1.00
	forest	865.00	0.53	0.50	0.00	0.00	1.00	1.00	1.00
	grassland	150.00	0.37	0.48	0.00	0.00	0.00	1.00	1.00
	harbour	510.00	0.28	0.45	0.00	0.00	0.00	1.00	1.00
	lake	355.00	0.20	0.40	0.00	0.00	0.00	0.00	1.00
	orchard	205.00	0.40	0.49	0.00	0.00	0.00	1.00	1.00
	residential	1055.00	0.54	0.50	0.00	0.00	1.00	1.00	1.00
	sparse shrub land	405.00	0.35	0.48	0.00	0.00	0.00	1.00	1.00
	sports land	160.00	0.62	0.49	0.00	0.00	1.00	1.00	1.00
	train station	260.00	0.22	0.41	0.00	0.00	0.00	0.00	1.00

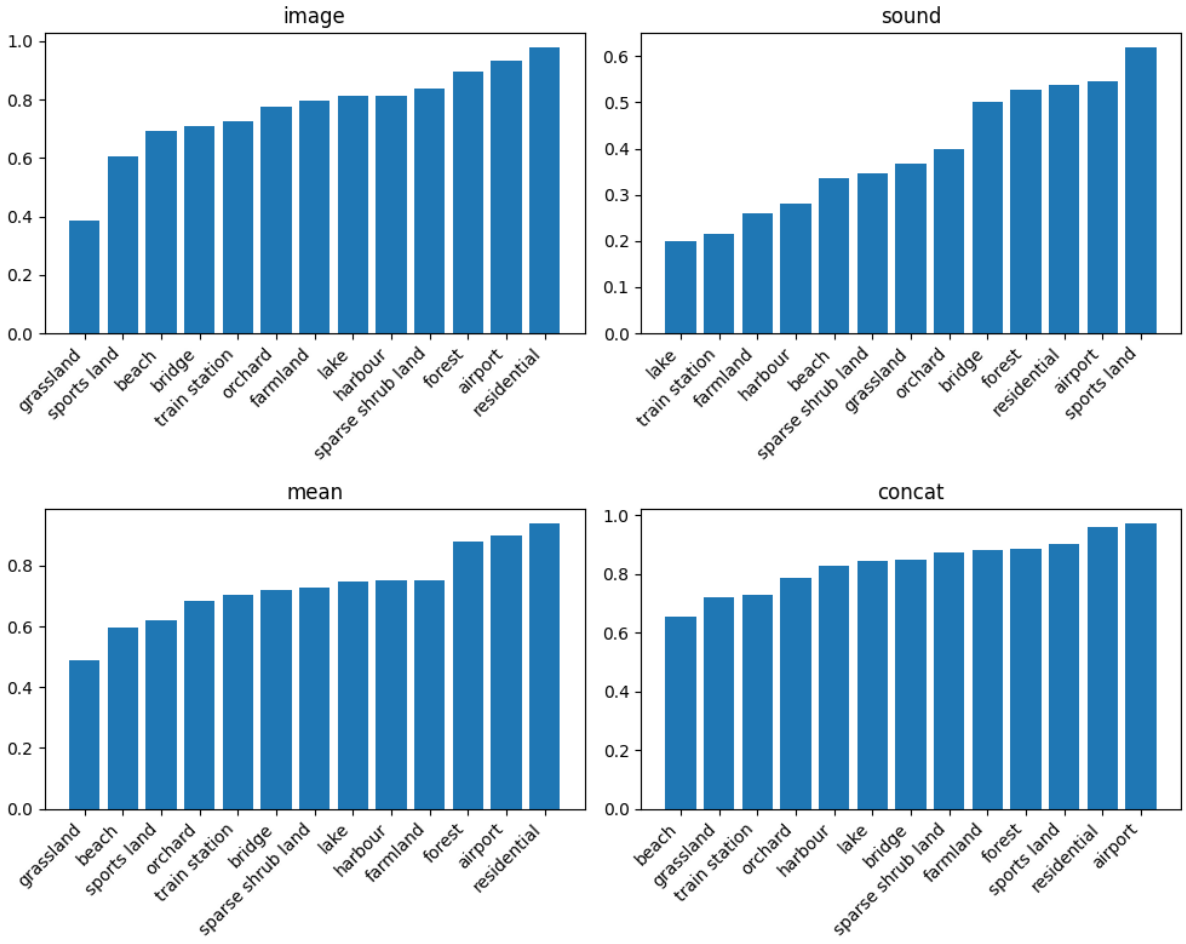
Fonte: Própria

Tabela 5 – Estatísticas por classes para ViT 16 patches para mean e concat.

mode	class name	count	mean	std	min	25%	50%	75%	max
concat	airport	185.00	0.97	0.16	0.00	1.00	1.00	1.00	1.00
	beach	215.00	0.66	0.48	0.00	0.00	1.00	1.00	1.00
	bridge	280.00	0.85	0.36	0.00	1.00	1.00	1.00	1.00
	farmland	430.00	0.88	0.33	0.00	1.00	1.00	1.00	1.00
	forest	865.00	0.89	0.32	0.00	1.00	1.00	1.00	1.00
	grassland	150.00	0.72	0.45	0.00	0.00	1.00	1.00	1.00
	harbour	510.00	0.83	0.38	0.00	1.00	1.00	1.00	1.00
	lake	355.00	0.84	0.37	0.00	1.00	1.00	1.00	1.00
	orchard	205.00	0.79	0.41	0.00	1.00	1.00	1.00	1.00
	residential	1055.00	0.96	0.20	0.00	1.00	1.00	1.00	1.00
	sparse shrub land	405.00	0.87	0.33	0.00	1.00	1.00	1.00	1.00
	sports land	160.00	0.90	0.30	0.00	1.00	1.00	1.00	1.00
	train station	260.00	0.73	0.45	0.00	0.00	1.00	1.00	1.00
mean	airport	185.00	0.90	0.30	0.00	1.00	1.00	1.00	1.00
	beach	215.00	0.60	0.49	0.00	0.00	1.00	1.00	1.00
	bridge	280.00	0.72	0.45	0.00	0.00	1.00	1.00	1.00
	farmland	430.00	0.75	0.43	0.00	1.00	1.00	1.00	1.00
	forest	865.00	0.88	0.33	0.00	1.00	1.00	1.00	1.00
	grassland	150.00	0.49	0.50	0.00	0.00	0.00	1.00	1.00
	harbour	510.00	0.75	0.43	0.00	1.00	1.00	1.00	1.00
	lake	355.00	0.75	0.44	0.00	0.00	1.00	1.00	1.00
	orchard	205.00	0.68	0.47	0.00	0.00	1.00	1.00	1.00
	residential	1055.00	0.94	0.24	0.00	1.00	1.00	1.00	1.00
	sparse shrub land	405.00	0.73	0.45	0.00	0.00	1.00	1.00	1.00
	sports land	160.00	0.62	0.49	0.00	0.00	1.00	1.00	1.00
	train station	260.00	0.70	0.46	0.00	0.00	1.00	1.00	1.00

Fonte: Própria

Figura 26 – Média de acertos por modo para ViT 16 patches.



Fonte: Própria

Tabela 6 – Estatísticas por classes para ViT 32 patches

mode	class name	count	mean	std	min	25%	50%	75%	max
image	airport	185.00	0.85	0.36	0.00	1.00	1.00	1.00	1.0
	beach	215.00	0.66	0.47	0.00	0.00	1.00	1.00	1.00
	bridge	280.00	0.83	0.37	0.00	1.00	1.00	1.00	1.00
	farmland	430.00	0.77	0.42	0.00	1.00	1.00	1.00	1.00
	forest	865.00	0.88	0.32	0.00	1.00	1.00	1.00	1.00
	grassland	150.00	0.59	0.49	0.00	0.00	1.00	1.00	1.00
	harbour	510.00	0.78	0.42	0.00	1.00	1.00	1.00	1.00
	lake	355.00	0.71	0.46	0.00	0.00	1.00	1.00	1.00
	orchard	205.00	0.71	0.45	0.00	0.00	1.00	1.00	1.00
	residential	1055.00	0.96	0.20	0.00	1.00	1.00	1.00	1.00
	sparse shrub land	405.00	0.86	0.35	0.00	1.00	1.00	1.00	1.00
	sports land	160.00	0.46	0.50	0.00	0.00	0.00	1.00	1.00
	train station	260.00	0.76	0.43	0.00	1.00	1.00	1.00	1.00
sound	airport	185.00	0.38	0.49	0.00	0.00	0.00	1.00	1.00
	beach	215.00	0.30	0.46	0.00	0.00	0.00	1.00	1.00
	bridge	280.00	0.48	0.50	0.00	0.00	0.00	1.00	1.00
	farmland	430.00	0.35	0.48	0.00	0.00	0.00	1.00	1.00
	forest	865.00	0.53	0.50	0.00	0.00	1.00	1.00	1.00
	grassland	150.00	0.35	0.48	0.00	0.00	0.00	1.00	1.00
	harbour	510.00	0.40	0.49	0.00	0.00	0.00	1.00	1.00
	lake	355.00	0.32	0.47	0.00	0.00	0.00	1.00	1.00
	orchard	205.00	0.45	0.50	0.00	0.00	0.00	1.00	1.00
	residential	1055.00	0.55	0.50	0.00	0.00	1.00	1.00	1.00
	sparse shrub land	405.00	0.33	0.47	0.00	0.00	0.00	1.00	1.00
	sports land	160.00	0.41	0.49	0.00	0.00	0.00	1.00	1.00
	train station	260.00	0.21	0.41	0.00	0.00	0.00	0.00	1.00

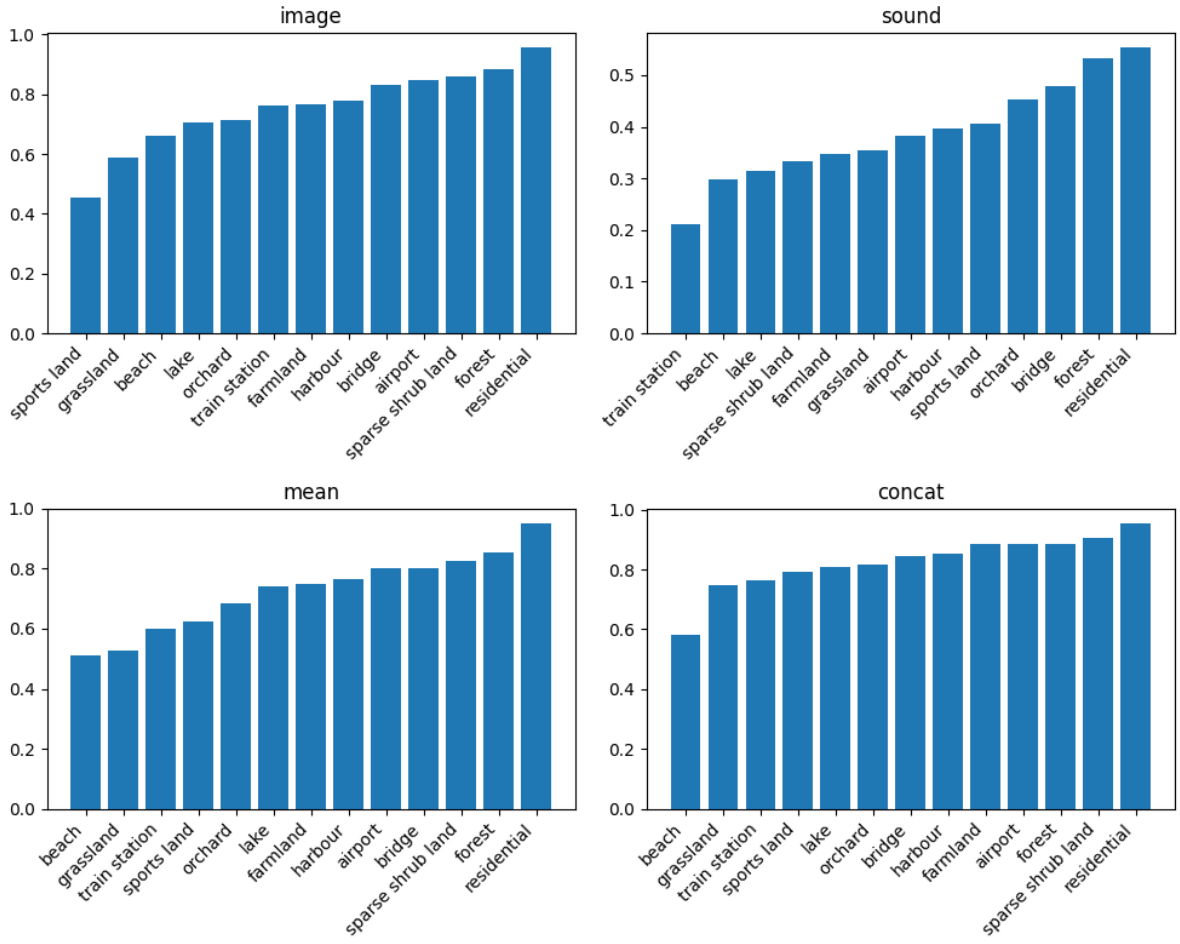
Fonte: Própria

Tabela 7 – Estatísticas por classes para ViT 32 patches

mode	class name	count	mean	std	min	25%	50%	75%	max
concat	airport	185.00	0.89	0.32	0.00	1.00	1.00	1.00	1.00
	beach	215.00	0.58	0.49	0.00	0.00	1.00	1.00	1.00
	bridge	280.00	0.85	0.36	0.00	1.00	1.00	1.00	1.00
	farmland	430.00	0.88	0.32	0.00	1.00	1.00	1.00	1.00
	forest	865.00	0.89	0.32	0.00	1.00	1.00	1.00	1.00
	grassland	150.00	0.75	0.44	0.00	0.25	1.00	1.00	1.00
	harbour	510.00	0.85	0.36	0.00	1.00	1.00	1.00	1.00
	lake	355.00	0.81	0.39	0.00	1.00	1.00	1.00	1.00
	orchard	205.00	0.81	0.39	0.00	1.00	1.00	1.00	1.00
	residential	1055.00	0.96	0.21	0.00	1.00	1.00	1.00	1.00
	sparse shrub land	405.00	0.90	0.30	0.00	1.00	1.00	1.00	1.00
	sports land	160.00	0.79	0.41	0.00	1.00	1.00	1.00	1.00
	train station	260.00	0.77	0.42	0.00	1.00	1.00	1.00	1.00
mean	airport	185.00	0.80	0.40	0.00	1.00	1.00	1.00	1.00
	beach	215.00	0.51	0.50	0.00	0.00	1.00	1.00	1.00
	bridge	280.00	0.80	0.40	0.00	1.00	1.00	1.00	1.00
	farmland	430.00	0.75	0.43	0.00	1.00	1.00	1.00	1.00
	forest	865.00	0.85	0.35	0.00	1.00	1.00	1.00	1.00
	grassland	150.00	0.53	0.50	0.00	0.00	1.00	1.00	1.00
	harbour	510.00	0.77	0.42	0.00	1.00	1.00	1.00	1.00
	lake	355.00	0.74	0.44	0.00	0.00	1.00	1.00	1.00
	orchard	205.00	0.68	0.47	0.00	0.00	1.00	1.00	1.00
	residential	1055.00	0.95	0.21	0.00	1.00	1.00	1.00	1.00
	sparse shrub land	405.00	0.83	0.38	0.00	1.00	1.00	1.00	1.00
	sports land	160.00	0.62	0.49	0.00	0.00	1.00	1.00	1.00
	train station	260.00	0.60	0.49	0.00	0.00	1.00	1.00	1.00

Fonte: Própria

Figura 27 – Média de acertos por modo para ViT 32 patches.



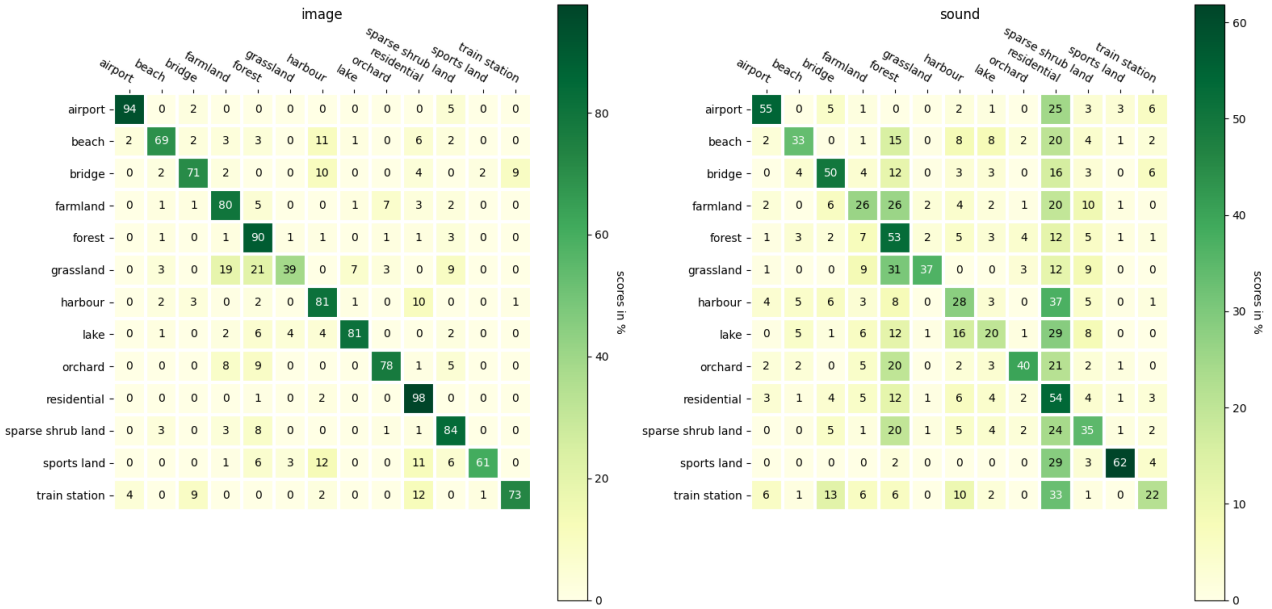
Fonte: Própria

Ao analisar o resultado por classes é visto que a quantidade de patches dos Vision Transformers modificam o desempenho de determinadas classes. Quando com 16 patches, as classes que apresentam mais detalhes seja no som ou na imagem tem desempenho melhor. Sports lands por ter um bom embedding de som apresenta um ótimo desempenho quando considerada com a imagem, farmland com bom embedding de imagem ao ser considerada com o som apresenta melhor desempenho também. Já com 32 patches, as melhores classes são as que apresentam características mais gerais e menos detalhadas como a sparse hub land que tem um melhor embedding de imagem e por isso ao ser considerada com seu som melhora a sua identificação, a classe bridge passa pelo mesmo pois seu embedding sonoro é mais característico e por isso ao ser considerado com a imagem tem melhor desempenho também.

6.3.2 Matrizes de confusão

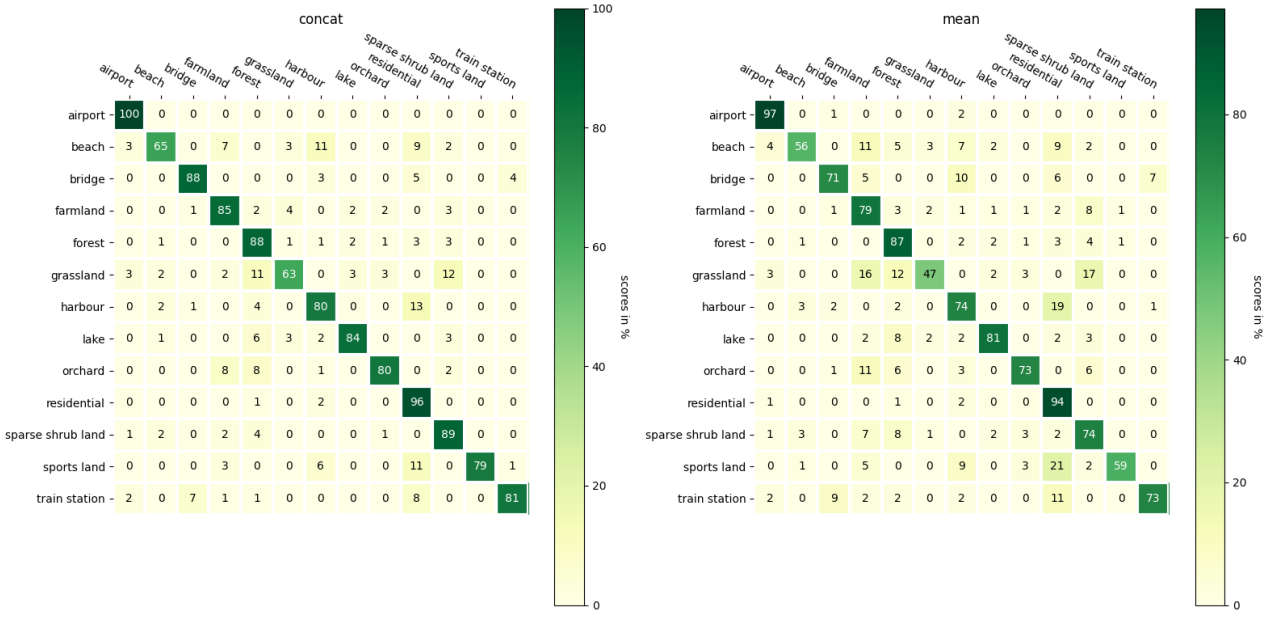
A matriz de confusão ilustra a quantidade de acertos por classe e tipo de embedding. Nela é possível observar as classes com mais acertos e erros e verificar os pontos de dificuldade dos modelos.

Figura 28 – Matriz de confusão para representações de image e sound para ViT 16 patches.



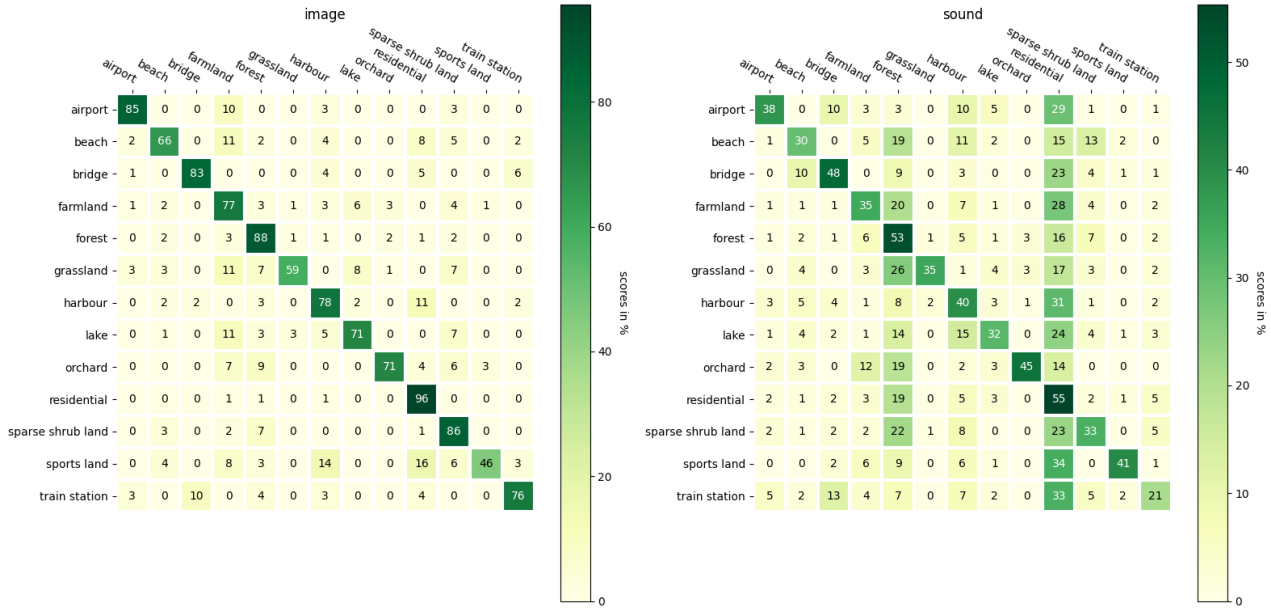
Fonte: Própria

Figura 29 – Matriz de confusão para representações de mean e concat para ViT 16 patches.



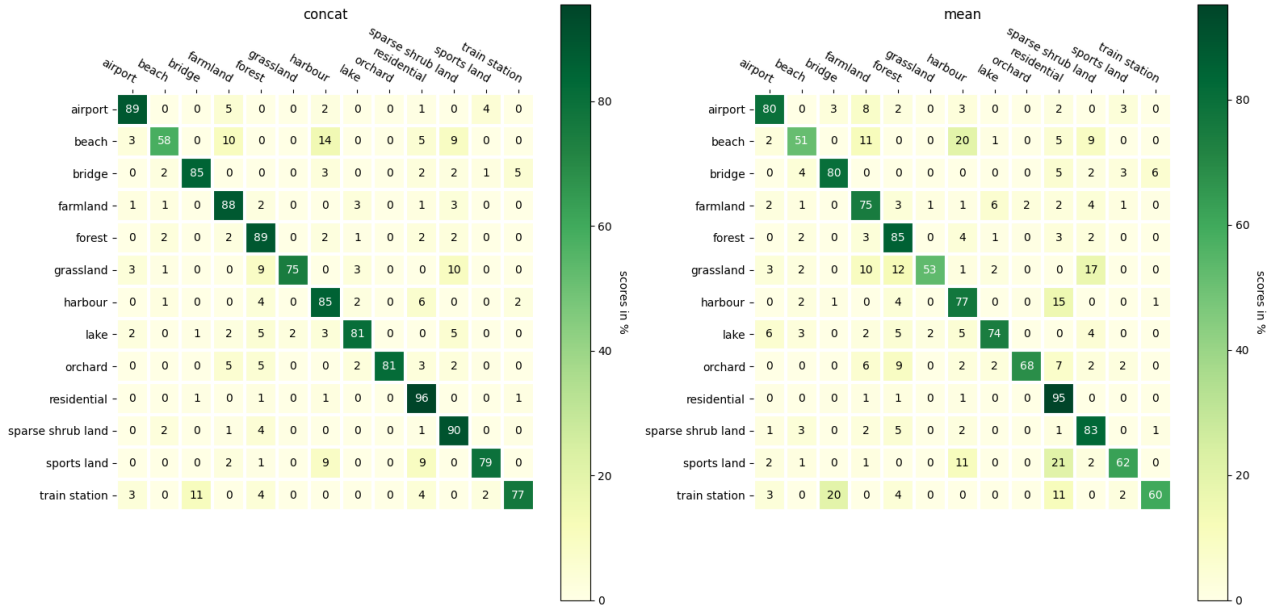
Fonte: Própria

Figura 30 – Matriz de confusão para representações de image e sound para ViT 32 patches.



Fonte: Própria

Figura 31 – Matriz de confusão para representações de mean e concat para ViT 32 patches.



Fonte: Própria

Visualizando as matrizes de confusão é visto melhor o que foi dito anteriormente, além de visualizar as classes com mais dificuldades de acerto, analisando apenas os embeddings de imagens: grassland, sports land e beach são as com mais erros (tanto com ViT

de 16 e 32 patches). Já para som vemos que as classes de forest e residential trouxeram mais dificuldades de identificação (tanto com ViT de 16 e 32 patches). Além disso, fica mais evidente também a influência de cada modalidade quando posta em conjunto (mean e concat), na matriz do ViT de 16 patches, é visto a melhora de desempenho das classes grassland,farmland e sports land, já na de 32 as classes vemos forest, harbour e orchard ter resultados melhores.

7 TRABALHOS FUTUROS

Realizar classificação de cenas aéreas com dados de som e imagem é uma nova área de pesquisa nos ramos de visão computacional, isso se dá pois, atualmente, com os recursos computacionais e dados disponíveis pela internet se tornou possível reunir informações a respeito dessa temática. Contudo, ainda existem desafios a serem seguidos:

- Falta de datasets em larga escala com dados de imagem e som em sensoriamento remoto: A escassez de datasets amplos para dados de imagem e som em sensoriamento remoto é evidente. Atualmente, apenas o ADVANCE serve como benchmark rotulado para algoritmos de machine/deep learning na classificação de cenas aéreas. O SoundingEarth destaca-se como o único dataset não rotulado que combina dados de imagem e som. Portanto, é crucial criar e desenvolver mais datasets desse tipo, incorporando novas classes e dados adicionais, para aprimorar os resultados obtidos com modelos de machine learning e deep learning.
- Aplicação de novas técnicas de aprendizagem de embeddings: É possível também realizar o desenvolvimento de pesquisa em novas maneiras de gerar os embeddings utilizando a abordagem do self-supervised learning com dados audiovisuais de sensoriamento remoto. Uma delas é a generativa onde se aplica *GANs (Generative Adversarial Networks)* que aprendem a diferenciar representações reais de falsas e por consequência aprender características importantes das imagens reais, logo, gerando boas representações. Para além disso, os transformers podem ser aplicados em diferentes arquiteturas, uma delas é o DINO (*self-DIstillation with NO labels*) onde as representações são geradas através de uma arquitetura de *student-teacher*, onde a *teacher network* tem seus pesos definidos como uma média móvel exponencial dos pesos da *student network*. Então, a *student network* é treinada para extrair os mesmos outputs da *teacher network* dado um par positivo (CARON et al., 2021). Por fim, outro ponto importante é realizar estudos na união dessas representações multimodais (*multimodal fusion*) como, por exemplo, aplicar o mecanismo de *attention* nesse caso. (MANDAL et al., 2024).
- Disponibilidade do código e modelo: Os modelos gerados com o vision transformers e código do projeto estão disponíveis em https://github.com/TalissaMoura /sounding_earth_with_vit. Com isso, espera-se que isso incentive futuras pesquisas com dados audiovisuais de sensoriamento remoto com modelos mais aprimorados ou que auxilie outras atividades que dependem desse tipo de dado.

8 CONSIDERAÇÕES FINAIS

Realizar classificação de cenas é uma das áreas da visão computacional que tem ganhado notoriedade recentemente por auxiliar em tarefas importantes como content based image retrieval, autonomous driving e smart content moderation. Para além disso, em sensoriamento remoto, esse problema também tem relevância por contribuir na tomada de decisões a respeito do nosso meio ambiente com monitoramento do clima e crescimento das cidades por exemplo.

Em relação a classificação de cenas utilizando dados de sensoriamento remoto, apesar da extensa pesquisa utilizando redes neurais convolucionais, técnicas novas como o self-supervised learning (SSL) e a ideia de construir arquiteturas sem depender de anotações de grandes datasets trouxeram melhores resultados nesse campo. Outra novidade recente, é a possibilidade de construir modelos multimodais, principalmente, com dados de imagem e áudio geolocalizados através dos datasets ADVANCE e SoundingEarth.

Nesse trabalho, pode-se perceber também que em conjunto com essas novas possibilidades, os vision transformers (ViT), paradigma mais recente na visão computacional, se mostrou com potencial para trazer novos avanços na resolução do problema de classificação de cena aéreas com dados audiovisuais. Através da técnica de SSL foi feito um pre-treino com o SoundingEarth gerando embeddings com ViT e utilizando a batch triplet loss para aproximar representações de áudio e imagem de um mesmo par e afastar representações de pares diferentes. Em seguida, elas foram aplicadas para a tarefa principal que é treinar um modelo linear para classificar as cenas aéreas do ADVANCE.

Dessa maneira, o ViT conseguiu gerar embeddings que conseguiram aprender características importantes das imagens e sons, obtendo precisão, recall e F1-Score de mais 80% para classificar cenas audiovisuais no ADVANCE. Quando utilizado apenas os embeddings de imagens, ele teve resultado superior a 80% e considerando apenas os embeddings de áudio, obteve-se mais de 40% nessas métricas.

Dito isso, apesar de um problema com potencial a ser resolvido utilizando deep learning, realizar a classificação de cenas audiovisuais de sensoriamento remoto apresenta alguns desafios como a falta de datasets em larga escala, principalmente, com dados multimodais, seja eles de imagens com áudio, imagens e texto ou com outros tipos de dados específicos a área como multi-spectral, hyperspectral e SAR. Outro ponto, é saber qual técnica apropriada de data augmentation aplicar nesses dados devido a suas propriedades específicas e, importante ressaltar que cenas aéreas mudam com tempo, dessa forma, outra dificuldade é construir modelos que aprendam características multi-temporais.

Sendo assim, o presente trabalho ilustrou os potenciais que o vision transformers e a técnica de SSL podem trazer para o problema de classificação de cena utilizando dados audiovisuais em sensoriamento remoto. Com a disponibilidade dos datasets, modelos e

código é possível que futuras pesquisas possam abordar modelos mais sofisticados com esses dados e também a partir dessas representações obter soluções para outras tarefas como segmentação de imagens, detecção de objetos, classificação do uso da terra e demais tarefas.

REFERÊNCIAS BIBLIOGRÁFICAS

AKYON, F. C.; TEMIZEL, A. ***Deep Architectures for Content Moderation and Movie Content Rating***. [S.l.]: arXiv, 2022.

ALZUBAIDI, L. et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. ***Journal of Big Data***, v. 8, n. 1, p. 53, mar. 2021. ISSN 2196-1115.

BALESTRIERO, R. et al. ***A Cookbook of Self-Supervised Learning***. [S.l.]: arXiv, 2023.

BALTRUŠAITIS, T.; AHUJA, C.; MORENCY, L.-P. ***Multimodal Machine Learning: A Survey and Taxonomy***. [S.l.]: arXiv, 2017.

BEGNINI, A. C. PREDICAO DE CLASSES SOCIAIS COM MODELOS DE APRENDIZADO PROFUNDO A PARTIR DE IMAGENS DE SATELITE E DADOS DE RENDA DO CENSO. 2023.

BERG, P.; PHAM, M.-T.; COURTY, N. Self-Supervised Learning for Scene Classification in Remote Sensing: Current State of the Art and Perspectives. ***Remote Sensing***, v. 14, n. 16, p. 3995, ago. 2022. ISSN 2072-4292.

BHATORE, S.; MOHAN, L.; REDDY, Y. R. Machine learning techniques for credit risk evaluation: A systematic literature review. ***Journal of Banking and Financial Technology***, v. 4, n. 1, p. 111–138, abr. 2020. ISSN 2524-7956, 2524-7964.

BYUN, S. et al. Road Traffic Monitoring from UAV Images Using Deep Learning Networks. ***Remote Sensing***, v. 13, n. 20, p. 4027, out. 2021. ISSN 2072-4292.

CARON, M. et al. ***Emerging Properties in Self-Supervised Vision Transformers***. [S.l.]: arXiv, 2021.

Carranza-García, M.; García-Gutiérrez, J.; RIQUELME, J. A Framework for Evaluating Land Use and Land Cover Classification Using Convolutional Neural Networks. ***Remote Sensing***, v. 11, n. 3, p. 274, jan. 2019. ISSN 2072-4292.

CASTELLUCCIO, M. et al. ***Land Use Classification in Remote Sensing Images by Convolutional Neural Networks***. [S.l.]: arXiv, 2015.

DEMARTY, C.-H. et al. VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation. ***Multimedia Tools and Applications***, v. 74, n. 17, p. 7379–7404, set. 2015. ISSN 1380-7501, 1573-7721.

DENG, J. et al. ImageNet: A Large-Scale Hierarchical Image Database. 2009–.

DOSOVITSKIY, A. et al. ***An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale***. [S.l.]: arXiv, 2021.

FACURE, M. ***Resolvendo CAPTCHAs com Redes Neurais Convolucionais***. 2017. Disponível em: ⟨https://matheusfacure.github.io/2017/03/12/cnn-captcha/⟩.

FENG, Q. et al. Multi-Temporal Unmanned Aerial Vehicle Remote Sensing for Vegetable Mapping Using an Attention-Based Recurrent Convolutional Neural Network. ***Remote Sensing***, v. 12, n. 10, p. 1668, maio 2020. ISSN 2072-4292.

FUJIYOSHI, H.; HIRAKAWA, T.; YAMASHITA, T. Deep learning-based image recognition for autonomous driving. ***IATSS Research***, v. 43, n. 4, p. 244–252, dez. 2019. ISSN 03861112.

GAD, A. F. ***Faster R-CNN explained for Object Detection Tasks***. Paperspace Blog, 2021. Disponível em: ⟨https://blog.paperspace.com/faster-r-cnn-explained-object-detection/⟩.

GANDHI, R. ***R-CNN, fast R-CNN, Faster R-CNN, YOLO - object detection algorithms***. Towards Data Science, 2018. Disponível em: ⟨https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e⟩.

GÉRON, A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2019.

GIRSHICK, R. et al. ***Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation***. [S.l.]: arXiv, 2014.

HE, K. et al. ***Deep Residual Learning for Image Recognition***. [S.l.]: arXiv, 2015.

HEIDLER, K. et al. ***Self-Supervised Audiovisual Representation Learning for Remote Sensing Data***. [S.l.]: arXiv, 2021.

HERRANZ, L.; JIANG, S.; LI, X. Scene recognition with CNNs: Objects, scales and dataset bias. In: ***2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***. [S.l.: s.n.], 2016. p. 571–579.

HOU, J. et al. Deep Quadruplet Appearance Learning for Vehicle Re-Identification. ***IEEE Transactions on Vehicular Technology***, v. 68, n. 9, p. 8512–8522, set. 2019. ISSN 0018-9545, 1939-9359.

HOWARD, A. G. et al. ***MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications***. [S.l.]: arXiv, 2017.

HU, D. et al. ***Cross-Task Transfer for Geotagged Audiovisual Aerial Scene Recognition***. [S.l.]: arXiv, 2020.

Jacques-Dumas, V. et al. Deep Learning-Based Extreme Heatwave Forecast. ***Frontiers in Climate***, v. 4, p. 789641, fev. 2022. ISSN 2624-9553.

JOSHI, A. et al. Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review. ***Remote Sensing***, v. 15, n. 8, p. 2014, abr. 2023. ISSN 2072-4292.

KARBALAIE, A.; ABTAHI, F.; SJÖSTRÖM, M. Event detection in surveillance videos: A review. ***Multimedia Tools and Applications***, v. 81, n. 24, p. 35463–35501, out. 2022. ISSN 1380-7501, 1573-7721.

LI, J. et al. Deep learning in multimodal remote sensing data fusion: A comprehensive review. ***International Journal of Applied Earth Observation and Geoinformation***, v. 112, p. 102926, ago. 2022. ISSN 15698432.

LI, Y. et al. A study on content-based video recommendation. In: ***2017 IEEE International Conference on Image Processing (ICIP)***. Beijing: IEEE, 2017. p. 4581–4585. ISBN 978-1-5090-2175-8.

LIU, S.; TIAN, G.; XU, Y. A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter. ***Neurocomputing***, v. 338, p. 191–206, abr. 2019. ISSN 09252312.

LORENTE, Ò.; RIERA, I.; RANA, A. ***Scene Understanding for Autonomous Driving***. [S.l.]: arXiv, 2021.

MANDAL, A. et al. ***Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection***. [S.l.]: arXiv, 2024.

MCCULLUM, N. ***Deep Learning Neural Networks explained in plain English***. freeCodeCamp.org, 2021. Disponível em: ⟨https://www.freecodecamp.org/news/deep-learning-neural-networks-explained-in-plain-english/⟩.

MEHMOOD, Z. et al. Content-Based Image Retrieval Based on Visual Words Fusion Versus Features Fusion of Local and Global Features. ***Arabian Journal for Science and Engineering***, v. 43, n. 12, p. 7265–7284, dez. 2018. ISSN 2193-567X, 2191-4281.

MEMON, J. et al. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). ***IEEE Access***, v. 8, p. 142642–142668, 2020. ISSN 2169-3536.

MINETTO, R. et al. Measuring Human and Economic Activity from Satellite Imagery to Support City-Scale Decision-Making during COVID-19 Pandemic. ***IEEE Transactions on Big Data***, v. 7, n. 1, p. 56–68, mar. 2021. ISSN 2332-7790, 2372-2096.

MORAUX, A. et al. Deep Learning for Precipitation Estimation from Satellite and Rain Gauges Measurements. ***Remote Sensing***, v. 11, n. 21, p. 2463, out. 2019. ISSN 2072-4292.

MOUTIK, O. et al. Convolutional Neural Networks or Vision Transformers: Who Will Win the Race for Action Recognitions in Visual Data? ***Sensors***, v. 23, n. 2, p. 734, jan. 2023. ISSN 1424-8220.

MUKHLIF, A. A.; Al-Khateeb, B.; MOHAMMED, M. A. Incorporating a Novel Dual Transfer Learning Approach for Medical Images. ***Sensors***, v. 23, n. 2, p. 570, jan. 2023. ISSN 1424-8220.

NOROOZI, M. et al. Boosting Self-Supervised Learning via Knowledge Transfer. In: ***2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition***. Salt Lake City, UT: IEEE, 2018. p. 9359–9367. ISBN 978-1-5386-6420-9.

OSCO, L. P. et al. A review on deep learning in UAV remote sensing. ***International Journal of Applied Earth Observation and Geoinformation***, v. 102, p. 102456, out. 2021. ISSN 15698432.

Prabhat et al. ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. ***Geoscientific Model Development***, v. 14, n. 1, p. 107–124, jan. 2021. ISSN 1991-9603.

QIANG, J. et al. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. ***IEEE Transactions on Knowledge and Data Engineering***, v. 34, n. 3, p. 1427–1445, mar. 2022. ISSN 1041-4347, 1558-2191, 2326-3865.

RAVENSCROFT, D. et al. Machine Learning Methods for Automatic Silent Speech Recognition Using a Wearable Graphene Strain Gauge Sensor. ***Sensors***, v. 22, n. 1, p. 299, dez. 2021. ISSN 1424-8220.

REDMON, J. et al. ***You Only Look Once: Unified, Real-Time Object Detection***. [S.l.]: arXiv, 2016.

REN, S. et al. ***Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks***. [S.l.]: arXiv, 2016.

RODRIGUEZ, R. ***6 machine learning steps explained for the business***. 2023. Disponível em: ⟨https://techbusinessguide.com/machine-learning-steps-seen-by-the-business/⟩.

SCHEIBENREIF, L. et al. Self-supervised Vision Transformers for Land-cover Segmentation and Classification. In: ***2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)***. New Orleans, LA, USA: IEEE, 2022. p. 1421–1430. ISBN 978-1-66548-739-9.

SHAFIQUE, A. et al. Deep Learning-Based Change Detection in Remote Sensing Images: A Review. ***Remote Sensing***, v. 14, n. 4, p. 871, fev. 2022. ISSN 2072-4292.

SHAH, D. ***Self-supervised learning and its applications***. 2023. Disponível em: ⟨https://neptune.ai/blog/self-supervised-learning⟩.

SIKIRIC, I. et al. Traffic Scene Classification on a Representation Budget. ***IEEE Transactions on Intelligent Transportation Systems***, v. 21, n. 1, p. 336–345, jan. 2020. ISSN 1524-9050, 1558-0016.

SIMONYAN, K.; ZISSERMAN, A. ***Very Deep Convolutional Networks for Large-Scale Image Recognition***. [S.l.]: arXiv, 2015.

SINGH, G. et al. Deep Learning in the Mapping of Agricultural Land Use Using Sentinel-2 Satellite Data. ***Geographies***, v. 2, n. 4, p. 691–700, nov. 2022. ISSN 2673-7086.

SREENU, G.; DURAI, M. A. S. Intelligent video surveillance: A review through deep learning techniques for crowd analysis. ***Journal of Big Data***, v. 6, n. 1, p. 48, dez. 2019. ISSN 2196-1115.

SRIVASTAVA, S.; Vargas-Muñoz, J. E.; TUIA, D. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. ***Remote Sensing of Environment***, v. 228, p. 129–143, jul. 2019. ISSN 00344257.

STOJNIĆ, V.; RISOJEVIĆ, V. ***Self-Supervised Learning of Remote Sensing Scene Representations Using Contrastive Multiview Coding***. [S.l.]: arXiv, 2021.

SZEGEDY, C. et al. ***Going Deeper with Convolutions***. [S.l.]: arXiv, 2014.

TAN, M.; LE, Q. V. ***EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks***. [S.l.]: arXiv, 2020.

TEKSANDS. ***Introduction to semi-supervised learning***. 2021. Disponível em: ⟨https://teksands.ai/blog/semi-supervised-learning⟩.

VOGEL, J.; SCHIELE, B. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. ***International Journal of Computer Vision***, v. 72, n. 2, p. 133–157, abr. 2007. ISSN 0920-5691, 1573-1405.

YEH, C. et al. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. ***Nature Communications***, v. 11, n. 1, p. 2583, maio 2020. ISSN 2041-1723.

YU, Z. et al. SegDetector: A Deep Learning Model for Detecting Small and Overlapping Damaged Buildings in Satellite Images. ***Remote Sensing***, v. 14, n. 23, p. 6136, dez. 2022. ISSN 2072-4292.

ZENG, D. et al. ***Deep Learning for Scene Classification: A Survey***. [S.l.]: arXiv, 2021.

ZHANG, C. et al. Detecting Large-Scale Urban Land Cover Changes from Very High Resolution Remote Sensing Images Using CNN-Based Classification. ***ISPRS International Journal of Geo-Information***, v. 8, n. 4, p. 189, abr. 2019. ISSN 2220-9964.

ZHANG, W.; YU, X.; HE, X. Learning Bidirectional Temporal Cues for Video-Based Person Re-Identification. ***IEEE Transactions on Circuits and Systems for Video Technology***, v. 28, n. 10, p. 2768–2776, out. 2018. ISSN 1051-8215, 1558-2205.

ZHOU, B. et al. Places: A 10 Million Image Database for Scene Recognition. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, v. 40, n. 6, p. 1452–1464, jun. 2018. ISSN 0162-8828, 2160-9292, 1939-3539.

ZHU, J. et al. ***Incorporating BERT into Neural Machine Translation***. [S.l.]: arXiv, 2020.

ZULFIQAR, M. et al. Deep Face Recognition for Biometric Authentication. In: ***2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)***. Swat, Pakistan: IEEE, 2019. p. 1–6. ISBN 978-1-72813-825-1.