Geração de descrição de obras de arte para deficientes visuais utilizando processamento de linguagem natural e visão computacional

Projeto final da disciplica de Inteligência Artificial

Pedro Henrique Barauna
UNIFESP
São José dos Campos, Brasil
phbarauna@unifesp.br

***Abstract*—Este trabalho pretende combinar processamento de linguagem natural com visão computacional para gerar descrições de obras de arte de Candido Portinari utilizando algoritmos de reconhecimento de objetos e geração de texto. Visando tornar as obras de arte mais acessíveis e compreensíveis para pessoas com deficiência visual. Os algoritmos de imagem e texto foram desenvolvidos para receber imagens de obras de arte e retornar uma descrição de cada obra. Também utiliza a API OpenAI para gerar um resumo da obra de arte e o módulo de voz Sapi5 da Microsoft para ler as descrições e o resumo em voz alta.**

***Index Terms*—Visão Computacional, Processamento de Linguagem Natural, OpenAI, Sapi5, Reconhecimento de Objetos, Geração de texto.**

I. Introdução

Uma das maiores características básicas de um ser humano é a visão. Um grande desafio para muitos dos deficientes visuais, é que eles são incapazes de ser completamente independentes. As pessoas com deficiência visual enfrentam problemas com atividades visuais, reconhecimento de objetos, desafios na identificação de itens e principalmente na apreciação e participação no mundo artístico tanto dentro, quanto fora, da internet.

Aplicações limitadas foram introduzidas no mercado para auxiliar pessoas com deficiência visual. No entanto, a ausência de uma aplicação voltada para a cultura e arte não é amplamente modernizada no ambiente web. Existem alguns aplicativos centrados em realidade aumentada, como, por exemplo, o MusA[8], desenvolvido por Dragan Ahmetovic, Kristian Keller, Cristian Bernareggi e Sergio Mascetti, na Universidade de Milão, porém, elas ainda não são o suficiente para a maioria das pessoas, pois é focado para museus e exposições e tem que ser previamente mapeado.

Inúmeras pesquisas são realizadas no domínio de reconhecimento de arte utilizando visão computacional e Deep Learning em diversos movimentos artísticos e tipos de arte, para o reconhecimento e bem como a criação de arte baseada nos mesmos. Porém, quase não se encontram projetos abertos que visam a criação de descrição dessas obras de arte ao mesmo tempo, de torná-la acessível para deficientes visuais.

O principal objetivo deste trabalho é apresentar uma plataforma de obras de arte na internet acessível para deficientes visuais, que gere descrições condizentes com a pintura escolhida, um resumo da obra e uma leitura em voz alta do que foi gerado. Utilizando modelos de codificação e decodificação de visão, que junta visão computacional e processamento de linguagem natural.

Em primeiro momento, será utilizado obras do artista mais famoso do Brasil Candido Portinari, do período da Arte Moderna, pintor de mais de quatro mil e quinhentas obras, que estão espalhadas por todo o mundo. As obras serão o conjunto de dados como treino do modelo para as descrições.

Conforme dito no "The Portinari Project: Science and Art team up together to help cultural projects"[1], Candido Portinari, viveu em um período muito significativo para a cultura moderna brasileira. Suas obras e suas interações com outros músicos, artistas, poetas, arquitetos, escritores, jornalistas, educadores e políticos, refletem a essência da preocupação estética, artística, cultural, social e política do Brasil no século XX. O Projeto Portinari, desenvolvido por Rosana Lanzelotte, Daniel Schwabe e João Candido Portinario, é o Catálogo Raisonné de todo o seu trabalho. Um banco de dados de imagens contendo também todas as informações do catálogo.

II. Conceitos Fundamentais

A. Candido Portinari

Candido Portinari nasceu em 29 de dezembro de 1903 em Brodowski, São Paulo. Sua infância humilde e suas experiências no meio rural influenciaram sua percepção das realidades sociais e moldaram sua arte. Portinari estudou na Academia Nacional de Belas Artes do Rio de Janeiro, onde estudou com importantes artistas brasileiros. Em sua obra, retrata o cotidiano dos trabalhadores rurais, pessoas simples e marginalizadas, destacando suas dificuldades e desigualdades.

A contribuição de Candido Portinari para a arte brasileira foi importante para a construção da cultura moderna do Brasil. Em suas pinturas, ele demonstra um compromisso social, usando sua habilidade artística para denunciar a injustiça e

a desigualdade vividas pelas camadas mais pobres da sociedade brasileira. Seus trabalhos mais conhecidos incluem ”O Lavrador de Café” (1939) e ”Guerra e Paz” (1956-1957), uma série de murais retratando a Segunda Guerra Mundial e a luta pela paz mundial.

Fig. 1. Pintura intitulada de ”O lavrador de café” de 1934

B. Machine Learning

O Machine Learning é um campo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que podem aprender e fazer previsões a partir de um conjunto de dados utilizando diversos métodos.

C. Redes neurais

As redes neurais são uma abordagem da inteligência artificial que ensina os computadores a processar dados de maneiras inspiradas no cérebro humano. É um processo de machine learning chamado Deep Learning que usa perceptrons interconectados, ou neurônios, em uma estrutura hierárquica, semelhante ao cérebro humano. As redes neurais criam um sistema adaptável que os modelos podem usar para aprender com seus erros e melhorar continuamente.

D. Perceptron e Multi-layer Perceptron

Em 1957, o psicólogo Frank Rosenblatt propôs ”O Perceptron: um autómato de percepção e reconhecimento” como uma classe de redes nervosas artificiais, incorporando aspectos do cérebro e dos receptores do sistema biológico biológicos[14]. Perceptron é a forma mais simples de configuração de uma rede neural artificial, cujo propósito focava em implementar um modelo computacional inspirado no cérebro humano. A simplicidade dele está associada à sua condição de ser constituída de apenas uma camada neural, tendo-se também somente um neurônio artificial. Ele pertence à arquitetura feedforward de camada única, pois o fluxo de informação em sua estrutura reside sempre no sentido da camada de entrada em direção à camada neural de saída, inexistindo-se qualquer tipo de realimentação de valores produzidos pelo seu único neurônio.

O Multi-layer Perceptron consiste numa camada de entrada, numa camada de saída e em camadas ocultas entre estas duas camadas [15]. Diferente do Perceptron onde existe apenas um neurônio de saída, a MLP pode relacionar diversas saídas de uma vez.

E. Processamento de linguagem natural

Processamento de Linguagem Natural (PLN) é uma área de Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais[13]. Ao longo dos anos, a PNL evoluiu de abordagens baseadas em regras para técnicas estatísticas e de machine learning, impulsionadas pela disponibilidade em excesso de dados linguísticos e poder de computação. A PNL abrange uma ampla gama de tarefas, incluindo, geração de linguagem, análise de sentimentos, tradução automática e sistemas de resposta a perguntas.

F. Visão computacional

Visão computacional é a ciência responsável pela visão de uma máquina, pela forma como um computador enxerga o meio à sua volta, extraindo informações significativas a partir de imagens capturadas por câmeras de vídeo, sensores, scanners, entre outros dispositivos.[12]

G. Modelo de encodificação e decodificação visual

Os modelos de encodificação e decodificação visual(Vision Encoder Decoder Models) são modelos de aprendizado de máquina usados no campo da visão computacional. Eles consistem em duas partes principais: um codificador e um decodificador. O codificador é responsável por extrair características relevantes de uma imagem, convertendo-a em um vetor de representação compacto sem perder informações. Por sua vez, o decodificador utiliza esse vetor para reconstruir a imagem original ou realizar outras tarefas, como classificação ou segmentação.

Um exemplo de aplicação, por sua vez, é a geração de descrição de imagens, utilizado no presente trabalho, em que o codificador é utilizado para codificar a imagem, e após isso o decodificador, um processador de linguagem natural, gera a descrição. Outro exemplo é o reconhecimento ótico de caracteres.

H. Fine-tuning

Fine-tuning é uma técnica de treinamento utilizada em aprendizado de máquina e em particular em modelos de linguagem. O fine-tuning envolve a etapa de treinamento adicional de um modelo pré-treinado em um conjunto de dados específico, geralmente menor, visando adaptar o modelo às características e nuances desse conjunto de dados específico. Em vez de treinar o modelo do zero, o fine-tuning aproveita o conhecimento prévio do modelo pré-treinado, treinado em um conjunto de dados grande e geral.

Fig. 2. Fluxo de um modelo encodificador decodificador

I. Cross-Entropy Loss

III. Trabalhos relacionados

Alguns artigos relacionados ao Candido Portinari, aplicação de Deep Learning para reconhecimento de obras de arte e traços e acessibilidade para deficientes visuais:

- The PORTINARI Project : Science and Art team up together to help cultural projects
  - O artigo conta sobre a história de Candido Portinari e do Projeto Portinari, suas fases, montagem de catálogo de obras, seu desenvolvimento e funcionamento nos dias de hoje.
- Automatic Tag Recommendation for Painting Artworks Using Diachronic Descriptions
  - Baseia-se na geração de categorias de forma automática para as obras de arte para melhor separá-las a partir de descrições de especialistas de arte brasileiros utilizando de um modelo de rede neural semi-superviosinado.
- A Study of Multi-Sensory Experience and Color Recognition in Visual Arts Appreciation of People with Visual Impairment
  - Este estudo conta sobre a possibilidade de experiências multi-sensoriais e reconhecimento de cores para deficentes visuais, e portanto, conseguirem apreciar exposições e museus de forma adequada.
- Object Recognition In Art Drawings: Tranfer Of A Neural Network
  - Consideram o problema de reconhecimento de objetos em desenhos e obras de arte com o objetivo de categorizar, pesquisar e organizar essas obras. Para tanto, é utilizado uma rede neural convolucional treinada com artes categorizadas.
- Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview
  - Apresenta uma visão geral de algumas das abordagens mais relevantes de Deep Learning, a partir dos recentes aumentos de dados de obras de arte e desenhos, para a extração e reconhecimento de padrões.
- Improving Object Detection in Art Images Using Only Style Transfer
  - Propõem e avaliam um processo diferente de treinamento de redes neurais para localizar pessoas em obras de arte, a partir de um conjunto de dados grande, criado especificamente para isso, modificado da base de dados do COCO, usando o estilo de transferência de AdaIn, e então, esse conjunto é utilizado em um modelo Fast R-CNN
- Large Scale Retrieval and Generation of Image Descriptions
  - Exploram métodos de para recuperar e gerar descrições em linguagem natural para imagens parecendo com o que pessoas reais escrevem, bem como condizer com o que é a imagem realmente. Para isso utilizaram de uma orientação aos dados utilizando técnicas de recupeção para buscas de descrições semelhantes com a imagem.
- MusA: Artwork Accessibility through Augmented Reality for People with Low Vision
  - Cita a dificuldade de pessoas deficientes visuais com pouca visão para ir a museus devido a problemas para encontrar as obras de arte. Para isso criaram um aplicativo inclusivo chamado MusA, que tem descrições dos museus, das obras de arte e também um mapa do museu.
- Accessibility, Usability and User Experience Design for Visually Impaired People: A Systematic Mapping Study
  - Estudo sobre como as tecnologias de comunicação e informação podem sem importantes para as pessoas deficientes visuais e como é ncessário que conceitos de acessibilidade, usabilidade e experiência do usuário são significantes no desenolvimento dessas tecnologias.
- Artificial Neural Networks and Deep Learning in the Visual Arts: a review
  - É efetuado uma análise da utilização de redes neurais e Deep Learning para artes visuais. Assim, é introduzido em como redes neurais conseguiram se aprofundar em diversar artes, como fotografia, modelagem 3D, artes conceituais, abstratas e entre outras.

No entanto, como pode ser observado nos trabalhos relacionados acima, de projetos de reconhecimento de arte, poucos trabalhos abordam com foco sobre deficientes visuais. Além disso, o MusA[8] mesmo que para deficientes visuais, foca em uma aplicação móvel em realidade aumentada e não um website aberto. Com o objetivo de preencher essas lacunas, esse projeto foi proposto.

IV. Objetivos

A seguir serão descritos os objetivos gerais e específicos

A. Gerais

O objetivo deste trabalho é o desenvolvimento de um modelo que gere descrições das obras de arte junto de uma plataforma web voltado para a acessbilidade para com deficientes visuais.

B. Específicos

- Tornar obras de arte acessíveis para pessoas deficientes visuais.
- Criar uma plataforma aberta e gratuita para a visualização acessível das obras de arte.

- Criar descrições condizentes com a obra de arte observada

V. Metodologia Experimental

Iremos adotar a seguinte metodologia:

- Importar a base de dados
- Importar um modelo visual como encodificador e um modelo pré treinado como decodificador
- Treinar o modelo
- Avaliar a geração de descrição
- Desenvolver a plataforma web
- Integrar o modelo a plataforma web

A rede será desenvolvido no Google Colab na linguagem Python, utilizando as bibliotecas PyTorch juntamente do Transformers, para a importação de modelos visuais e modelos pré-treinados para a criação da arquitetura visual encoder-decoder. O modelo de voz será da biblioteca Pyttsx3 e escolhida a voz padrão feminina da Microsoft "Sapi5". A voz não funciona em ambiente Linux, portanto, tornando necessário o sistema operacional do Windows para rodá-la. Treinamento e teste do modelo será feito a partir das artes de Candido Portinari e algumas obras de período da arte moderna. Além destas, também haverão bibliotecas auxiliares como NumPy, Matplot e entre outras. O design foi criado na plataforma Figma, para melhor prototipação de animações e será desenvolvida com HTML, CSS e Javascript.

A. Base de Dados

A base de dados é composta de 70 obras de arte do artista e pintor Candido Portinari, retiradas do Projeto Portinari. Suas histórias, descrições e data de criação, conseguirão ser acessadas da mesma forma pelo website e serão utilizadas para melhorar as descrições ao passar do projeto. Também foram selecionadas outras obras modernistas pós-impressionistas retiradas do WikiArt, enciclopédia artística de obras, artistas e movimentos.

B. Modelos

Foram utilizados inicialmente, modelos pré-treinados da biblioteca HuggingFace, um grande repositório de modelos de datasets para o uso na licença devida por outros usuários. Os modelos escolhidos para teste foram "nlpconnect/GPT-2", "abdou/Swim-Base GPT-2" e "bipim/GPT-2". Todos utilizam como gerador de texto o GPT-2, porém todos utilizam de diferentes modelos de visão de encodificação e decodificação da imagem.

Nlpconnect/GPT-2: Este modelo tem como decodificador o GPT-2 e como codificador um modelo de transformação de visão(ViT) próprio do HuggingFace. Utiliza em sua base o PyTorch para criação do modelo encodificador-decodificador na totalidade. Abdou/Swim-Base GPT-2: Este modelo tem como decodificador o GPT-2 e como codificador um modelo de transformação de visão(ViT) da microsoft, o Swin Transformer. Utiliza também em sua base o PyTorch para criação do modelo encodificador-decodificador e foi pré-treinado com 60Bipim/GPT-2: Este modelo tem como decodificador o GPT-2 e como codificador um modelo de transformação de visão(ViT) próprio do HuggingFace também. Porém, ele é menos pré-treinado e mais adaptável a outros treinos.

C. Treino de modelo

Para o treino dos modelos, foi utilizado 46 obras de arte com a criação de descrições condizentes para cada obra e, assim, será realizado o Fine-Tuning dos modelos, descongelando-os e tornando-os mais específicos para obras de arte modernas. Para a otimização, foi utilizado a função de custo de Cross-Entropy, que é considerada melhor para o reconhecimento de novas classes, como dito por Elliott Gordon-Rodriguez, Gabriel Loaiza-Ganem, Geoff Pleiss, John Patrick Cunningham no artigo "Uses and Abuses of the Cross-Entropy Loss: Case Studies in Modern Deep Learning"[21]

D. Plataforma web

A criação da plataforma web é visada para pessoas com visibilidade reduzida até deficientes visuais, assim, criando um ambiente inclusivo para as pessoas.

Todo o design da plataforma foi criado com uma profissional da área de design de experiência do usuário, focando nos conceitos de acessibilidade encontrados no paper [9] do João Ricardo dos S. Rosa e da Natasha Malveira C. Valentim. e no capítulo [22] para criação de materiais acessíveis a pessoas com deficiência visual.

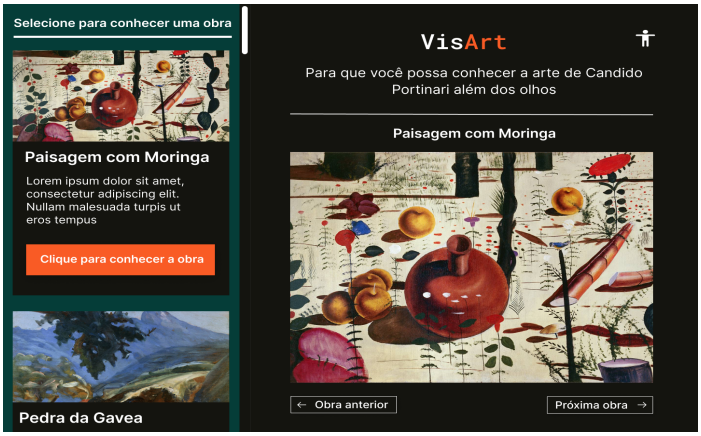

Fig. 3. Design dark-mode da plataforma

E. Protocolo de validação

Para validar as descrições, foi utilizado um conjunto de palavras genéricas, mas com contexto das obras de arte da base de dados. A partir disso, é montada uma estrutura de dados de árvore, assim, um dicionário. Então, com as descrições geradas, é separada cada palavra com mais de 4 letras e verificando se existe ela na árvore e trazendo a porcentagem de palavras-chave na descrição. Posteriormente, será possível gerar uma árvore de validação para cada obra e então escolher descrições mais apropriadas com base nas palavras-chave. Também, foi realizada validação manual, sendo realizada a leitura das descrições geradas e verificando sua semântica e o entendimento realizado pelo gerador de descrições nas obras.

F. Diagrama de blocos e fluxo

A obra de arte é o início do fluxo, ou seja, será escolhida a obra, onde, após isso, será cortada em 5 partes e gerada a descrição para cada corte, traduzida e lida pelo Sapi5, caso não seja condizente, ela é gerada novamente. As descrições são validadas conforme o dicionário criado, citado no protocolo de validação, porém, apenas com foco de detectar palavras-chave não utilizadas e adicioná-las para as próximas validações.

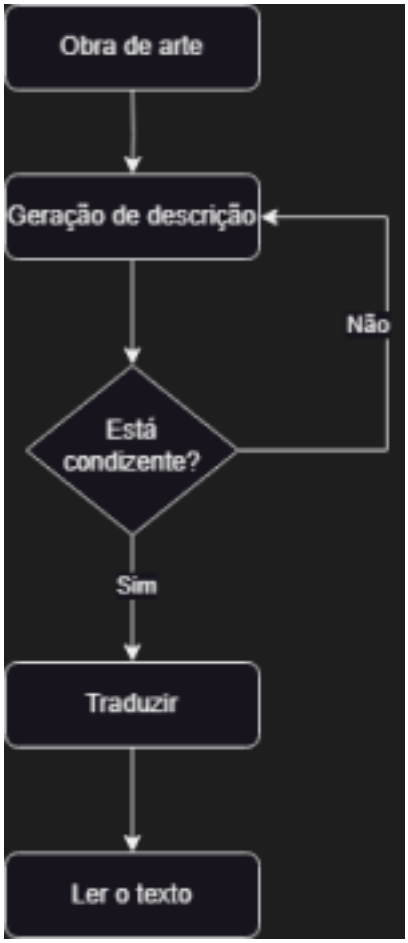

Fig. 4. Diagrama de fluxo do algoritmo

VI. Resultados e Discussões

Utilizando das 70 obras de artes do Projeto Portinari, a que gerou menos palavras presentes na árvore ontológica foi o de Bipim. Como dito na seção de ”Modelos”, ele é menos treinado comparado aos outros e para a utilização em obras de arte acaba sendo inviável, pois será necessário muito refinamento e mais treinos para torná-lo apresentável. O de Abdou foi o segundo melhor em geração de palavras-chave, ele conseguiu atribuir quase o dobro de Bipim, porém, mesmo aumentando seu treino e refinamento, ele não demonstrou melhora nas descrições. O melhor modelo que se adaptou ao treino e conseguiu utilizar de mais palavras-chave que descreviam a obra foi o Nlpconnect. Ele é um modelo pré-treinado mais denso, mas que foi possível adicionar mais camadas de aprendizado.

Modelos Pré-treinados	Média de palavras-chave acertadas
nlpconnect/GPT-2	65.21
abdou/Swim-Base GPT-2	46.03
bipim/GPT-2	27.17

Fig. 5. Média de palavras-chaves acertadas por cada modelo

Também foram realizados testes manuais, ou seja, todas as leituras foram lidas por uma profissional da área. E os resultados prosseguiram da mesma forma, sendo Bipim, o menos condizente, o Abdou pouco mais condizente e o Nlpconnect como o melhor condizente, com a maioria das obras. Assim, o modelo de Nlpconnect foi o que melhor performou diante dos testes e será utilizado como modelo final para a geração de descrições das obras de arte. Ele será mais treinado e refinado para melhores resultados posteriores.

Para mais testes, foram recolhidas mais imagens modernistas e a partir delas, foi realizada a geração de descrição para averiguar a efetividade do dicionário de palavras. Então, foi apresentado que mesmo sendo imagens diferentes, há o uso repetitivo de palavras-chave como, por exemplo, ”man, woman” com certa efetividade.

Mas, mesmo o Nlpconnect sendo o melhor modelo, o treino realizado utilizando a Cross-Entropy Loss foi efetivo, porém de forma limitada. Ainda não são utilizadas palavras que seriam o foco para obras de arte, como conceitos de cores e iluminação. A plataforma utilizada para a realização do treino aparenta ser de pouca disponibilidade de memória e processamento para os novos layers que serão criados por cima do modelo pré-treinado e para tal tarefa, demandaria, possivelmente em uma maior quantidade.

Junto disso, foi visto que, o modelo não aprendeu contextos da imagem, ou seja, ao tentar dividi-la em quadrantes, não foi efetiva a geração de descrições desse quadrante, sendo então, necessário descrições mais longas para avaliar toda a obra, ou uma diminuição do número de quadrantes. Porém, conforme aumentamos o número de sentenças e classes para gerar, mais dispersa vai ficando a descrição e menos assertiva ela fica, utilizando de repetições de sinônimos para entendê-la. Portanto, é necessário um meio-termo para a descrição sair o mais condizente possível.

A plataforma desenvolvida a primeiro momento, contava com os quadrantes, mas visto a necessidade de tornar mais assertivo, será verificada a melhor opção entre a quantidade quadrantes ou uma descrição maior da obra toda.

VII. Conclusão

Foi percebido então, que a criação de um projeto em inteligência artificial de obras de arte voltado para a acessibilidade de pessoas deficientes visuais é de extrema importância e necessidade, visto que para este contexto, não existem pesquisas o suficiente sobre o tema. Os projetos que tangem o tema, baseiam-se na criação de um aplicativo voltado

para ser presencialmente utilizado em museus artísticos ou no reconhecimento em tempo real de objetos, mas que não contemplam a visualização de obras de arte.

Utilizar da arte moderna como movimento escolhido para o treino e teste do modelo foi difícil e complexo, já que existem certos abstracionismos que não conseguem ser identificados pelo modelo, da forma correta, bem como significados ambíguos em formas das obras. Mas, foi possível mostrar a importância de Candido Portinari, artista brasileiro, no movimento modernista.

O modelo desenvolvido em questão, conseguiu ser parcialmente assertivo para o todo do projeto, mas devido a certas limitações, como o entendimento do contexto da obra, seria necessário mais desenvolvimento e poder de processamento, para construir mais camadas focadas para a detecção de classes dentro das obras de arte.

E, dentro disso, a partir da criação da plataforma, será possível demonstrar, mesmo que seja parcialmente, uma obra de arte para pessoas deficientes visuais de forma livre e gratuita e também demonstrando a importância da arte perante a acessibilidade na internet.

VIII. Trabalhos futuros

Como dito anteriormente, existe a necessidade de melhorar a análise do modelo, com mais treino, teste e uma melhor forma de avaliação do que foi gerado. Com isso, será possível obter mais assertividade e contexto das descrições com a possível divisão da obra em quadrantes, que era o intuito inicial do projeto.

IX. Agradecimentos

Os dados da base foram retirados do Projeto Portinari, com agradecimentos aos atuadores do projeto. Também, os modelos pré-treinados utilizados foram fornecidos por desenvolvedores de software livre, que disponibilizaram os repositórios de forma gratuita e aberta e que foram de suma importância ao projeto.

References

[1] Lanzelotte, Rosana Marques, M. Schwabe, Daniel Penna, M. Portinari, Joao Ruiz, F.. (1993). The Portinari Project - Science and Art Team Up Together to Help Cultural Projects.. 146-157.

[2] Zuin, G., Veloso, A., Portinari, J. C., Ziviani, N. (2020). Automatic Tag Recommendation for Painting Artworks Using Diachronic Descriptions. 2020 International Joint Conference on Neural Networks (IJCNN)

[3] Cho, J.D. A Study of Multi-Sensory Experience and Color Recognition in Visual Arts Appreciation of People with Visual Impairment. Electronics 2021, 10, 470

[4] Yin, R., Monson, E., Honig, E., Daubechies, I., Maggioni, M. (2016). Object recognition in art drawings: Transfer of a neural network. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[5] Castellano, G., Vessio, G. Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. Neural Comput Applic 33, 12263–12282 (2021).

[6] D. Kadish, S. Risi and A. S. Lovlie, Improving Object Detection in Art Images Using Only Style Transfer, 2021.

[7] Ordonez, V., Han, X., Kuznetsova, P., Kulkarni, G., Mitchell, M., Yamaguchi, K., … Berg, T. L. (2015). Large Scale Retrieval and Generation of Image Descriptions. International Journal of Computer Vision, 119(1), 46–59.

[8] D. Ahmetovic, K. Keller, C. Bernareggi and S. Mascetti, MusA: Artwork Accessibility through Augmented Reality for People with Low Vision.

[9] Rosa, J. R. dos S., Valentim, N. M. C. (2020). Accessibility, usability and user experience design for visually impaired people. Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems.

[10] Santos, I., Castro, L., Rodriguez-Fernandez, N., Torrente-Patiño, Á., Carballal, A. (2021). Artificial Neural Networks and Deep Learning in the Visual Arts: a review. Neural Computing and Applications, 33(1), 121–157.

[11] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[12] de Milano, D., Honorato, L. B. (2010). Visao computacional. Faculdade de Tecnologia, Universidade Estadual de Campinas.

[13] Vieira, Renata, and Lucelene Lopes. "Processamento de linguagem natural e o tratamento computacional de linguagens científicas." Em corpora (2010): 183.

[14] Kanal, Laveen N. "Perceptron." Encyclopedia of Computer Science. 2003. 1383-1385.

[15] Ramchoun, Hassan, et al. "Multilayer perceptron: Architecture optimization and training." (2016).

[16] Hugging Face: Repositório online de datasets e modelos. Disponível em: Hugging Face

[17] Hugging Face: Vision Encoder Decoder Model. Disponível em: Hugging Face - Vision Encoder Decoder

[18] Hugging Face: nlpconnect/vit-gpt2-image-captioning. Disponível em: Hugging Face - nlpconnect/vit-gpt2-image-captioning

[19] Hugging face: bipin/image-caption-generator. Disponível em: Hugging Face - bipin/image-caption-generator

[20] Hugging Face: Abdou/vit-swin-base-224-gpt2-image-captioning. Disponível em: Hugging Face - Abdou/vit-swin-base-224-gpt2-image-captioning

[21] Gordon-Rodriguez, E., Loaiza-Ganem, G., Pleiss, G. amp; Cunningham, J.P.. (2020). Uses and Abuses of the Cross-Entropy Loss: Case Studies in Modern Deep Learning.

[22] Simões, P. Aliana, Frizzera, S. C. Ana, Koehler, D. Andressa, Sondermann, C. V. Danielli. Cap. 5: O leitor de tela e a criação de materiais digitais acessíveis a pessoas com deficiência visual.