Universidade de São Paulo
Escola Politécnica
Programa de Educação Continuada em Engenharia
Especialização em Inteligência Artificial

Vinícius Grandolpho Selestrim

Classificação de movimentos utilizando modelos fundacionais

São Paulo
2023

VINÍCIUS GRANDOLPHO SELESTRIM

Classificação de movimentos utilizando modelos fundacionais

— Versão Original —

Monografia apresentada ao Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para conclusão do curso de Especialização em Inteligência Artificial.

Orientador: Prof. Dr. Eduardo Lobo Lustosa Cabral

São Paulo
2023

Catalogação-na-publicação

Selestrim, Vinícius Grandolpho
Classificação de movimentos utilizando modelos fundacionais/ V.Selestrim – São Paulo, 2023.
150p.

Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1. Visão Computacional 2. Modelo Fundacional 3. Classificador De Movimentos.
I. Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia. II.t.

Para Bruna, que sempre me dá forças para continuar.

Agradecimentos

Agradeço à minha família que sempre apoiou minha busca pelo conhecimento. Agradeço aos professores do curso, em especial meu orientador Eduardo Lobo Lustosa Cabral que esteve sempre disponível a sanar quaisquer dúvidas, além dos professores Thiago de Castro Martins e Larissa Driemeier por todo conhecimento compartilhado. Finalmente, agradeço ao professor Marcelo Knörich Zuffo, pois sem sua recomendação nada disso teria sido possível.

Sumário

Resumo

SELESTRIM, V.*Classificação de movimentos utilizando modelos fundacionais.* 2023. Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia. Universidade de São Paulo, São Paulo, 2023.

A visão computacional se tornou ferramenta fundamental na vida das sociedades modernas. Seja para auxiliar as pessoas em tarefas complexas e exaustivas, seja para identificar os diversos objetos ao redor de um carro autônomo, a classificação de objetos mudou a forma que as máquinas interagem com os humanos. Assim como diversas soluções foram originadas com o advento das redes neurais para a classificação de imagens, um novo leque de possibilidades se abre com a proposta de um novo método para a classificação de movimentos. Com o uso de um modelo fundacional para extração de características é possível obter particularidades de imagens, levando em consideração aspectos espaciais da mesma. Um conjunto de imagens pode ser uma sequência ordenada originada a partir de um video. Tal conjunto de imagens, ao ter suas características extraídas, gera uma série temporal única para tal video. Ao treinar um modelo de inteligência artificial alimentado por essas séries temporais é possível criar um classificador de movimentos. Diferentemente de outros trabalhos cujo foco é a classificação de movimentos de pessoas ou de objetos específicos, esta abordagem permite que o mesmo modelo seja treinado para quaisquer tipos de movimentos, seja de objetos animados ou inanimados. Testes preliminares mostraram que trata-se de uma técnica promissora e, além de permitir novas soluções, pode trazer avanços em áreas que necessitavam de uma nova abordagem para prosseguirem seu desenvolvimento. De carros efetuando manobras proibidas a gestos realizados com as mãos, a técnica aqui proposta é capaz de classificar qualquer tipo de movimento, o que a torna especialmente relevante.

Palavras-chave: Visão Computacional. Modelo Fundacional. Classificador De Movimentos.

Abstract

SELESTRIM, V.*Motion classifier using foundation models.* 2023. Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia. University of São Paulo, São Paulo, Brazil. 2023.

Computer vision has become a fundamental tool in the lives of modern societies. Whether assisting individuals in complex and exhaustive tasks or identifying various objects around an autonomous vehicle, object classification has transformed the way machines interact with humans. Just as various solutions emerged with the advent of neural networks for image classification, a new range of possibilities arises with the proposal of a novel method for motion classification. By using a foundation model for feature extraction, it is possible to obtain specific details from images, taking into account their spatial aspects. A set of images can be an ordered sequence derived from a video. This set of images, once its features are extracted, generates a unique time-series for that video. Training an artificial intelligence model feeded by these time-series enables the creation of a motion classifier. Unlike other papers focused on the classification of people movements or specific objects, this approach allows the same model to be trained for any types of movements, whether animated or inanimate objects. Preliminary tests have shown that it is a promising technique and, in addition to enabling new solutions, it may bring advancements in areas that needed a new approach to continue their development. From cars performing prohibited maneuvers to hand gestures, the proposed technique can classify any type of movement, making it particularly relevant.

Keywords: Computer Vision. Foundation Model. Motion Classifier,

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A visão computacional determinou um ponto de inflexão na forma que as máquinas interagem com os seres humanos. Desde o momento em que a rede AlexNet competiu e venceu o ImageNet Large Scale Visual Recognition Challenge, em Setembro de 2012, os avanços desta área foram significativos. A possibilidade de identificar objetos em imagens com alta precisão possibilitou a criação de diversas soluções que até então não eram possíveis, ou eram feitas de forma imprecisa por seres humanos. Por mais que a tarefa de classificação de objetos em imagens tenha sido considerada concluída devido ao avanço dos modelos de redes neurais artificiais, a classificação de movimentos ainda não atingiu sua maturidade. Visto como um próximo passo da classificação de objetos, a classificação de movimentos possibilita novas soluções. Neste trabalho é apresentada uma nova abordagem ao tema, que mostra-se promissora em seus resultados.

A classificação de movimentos através da visão computacional possibilita a criação de soluções que podem ser exaustivas aos olhos humanos. Neste trabalho é feita uma abordagem onde qualquer movimento pode ser classificado, seja este realizado por uma pessoa, objeto, animal, ou mesmo causado a partir de questões naturais. Desde que exista um conjunto de dados com videos suficientes do movimento que se quer classificar, qualquer movimento pode ser identificado. Esta abordagem se diferencia das abordagens tradicionais de classificação de movimento, onde normalmente o foco do estudo são movimentos do corpo humano (classificação de poses), ou movimentos específicos das mãos ou de um objeto. A partir deste estudo, avanços em diversas áreas podem ser explorados, como na identificação de manobras proibidas no trânsito, identificação de anomalias em funcionamento de máquinas, prevenção de desastres naturais, etc.

A abordagem aqui proposta se torna factível devido ao surgimento dos chamados modelos fundacionais ou modelos base da visão computacional. Os modelos fundacionais podem ser entendidos como modelos criados a partir de uma quantidade de dados especialmente grande (conjunto de dados massivos), que podem ser usados para qualquer

tipo de tarefa, uma vez que não possuem uma única saída possível. Até então, as redes pré-treinadas perdiam informações importantes das imagens que compõem os videos cujo movimento se quer classificar, principalmente no que diz respeito à posição espacial de objetos em movimento. A extração de características usando modelos fundacionais resolve este problema, e torna esta abordagem possível de ser implementada.

O objetivo deste trabalho é criar um novo método com o qual as pessoas podem gravar videos com movimentos quaisquer, de forma que os movimentos sejam classificados com a finalidade de executar alguma tarefa. Este trabalho consiste no início de um desenvolvimento que posteriormente pode englobar aplicativos com diversas funções, tais como, controle de periféricos em uma casa (volume de uma televisão, acender e apagar de luzes) e outras funções que pessoas com mobilidades reduzidas podem se beneficiar. Atualmente existem assistentes por voz que podem executar funções similares, porém o método apresentado neste trabalho agrega mais uma ferramenta para produtos voltados à acessibilidade. A principal vantagem desse método é permitir o seu uso independente do nível de ruído no local, já que funciona em ambientes com muito barulho ou em ambientes onde não é permitido fazer qualquer tipo de ruído.

Os exemplos utilizados neste trabalho são compostos de movimentos feitos a partir de gestos com as mãos. Uma vez que dois dos movimentos classificados são relativamente parecidos, é possível observar a robustez desta abordagem. Nos próximos capítulos é mostrada a forma que esta abordagem foi implementada, os modelos utilizados para atingir os objetivos propostos e os resultados, além de discutir possíveis melhorias para futuras implementações.

Capítulo 2

Revisão da literatura

O tema de classificação de movimentos já foi visitado por diversos autores. Alguns trabalhos se destacam por sua precisão nos resultados. O trabalho de Ahmed et al. (2020) mostra um método eficiente de classificar movimentos humanos com uma rede relativamente simples, obtendo resultados relevantes no que diz respeito à sua exatidão. O trabalho em questão propõe dividir o video que se quer classificar em imagens ordenadas (quadros do video), e realizar uma operação de diferença entre as imagens. Novas imagens são geradas contendo as diferenças e estas imagens passam por uma rede neural convolucional, com a finalidade de extrair as características das diferenças. Após duas camadas convolucionais e uma de "max pooling", o resultado é colocado em uma rede com camadas densas (totalmente conectadas) com duas saídas (considerando um treinamento com duas classes). Esta abordagem se mostrou pertinente para os conjuntos de dados utilizados pelo autor, obtendo taxas de acerto elevedas. O problema encontrado pelo autor se deu em videos com quadros sem algum tipo de movimento, o que ele chama de fundo constante. Foi necessário utilizar uma técnica para eliminar quadros sem movimentos, uma vez que a imagem resultante da diferença entre quadros sem movimentos poderia alterar o funcionamento da rede.

O fato de haverem quadros sem qualquer tipo de movimento não é um problema para o trabalho proposto por Xing, Di Caterina e Soraghan (2020). Nesta abordagem, os autores convertem as imagens que compõem os videos em valores proporcionais a pulsos elétricos, o que chamam de "spikes", utilizando sensores de visão dinâmica (DVS) no lugar de câmeras convencionais. Enquanto este tipo de sensor imita o funcionamento da retina humana, a rede proposta pelos autores imita o funcionamento do cérebro humano. Além de camadas convolucionais, a rede proposta conta com camadas recorrentes, que introduzem memória e a habilita para trabalhar com dados sequênciais. Como resultado, a rede SCRNN ("Spiking Convolutional Recurrent Neural Network") obteve uma exatidão de 96.59% para classificação de 10 diferentes movimentos do "DVS gesture dataset"da

IBM, conjunto de dados gerado para ser usado no trabalho de Amir et al. (2017). Trata-se de um resultado notável para um conjunto de dados de tamanha complexidade. Por mais que esta abordagem aparenta resolver a tarefa de classificação de movimentos quaisquer, sensores de visão dinâmica são dispositivos caros e restritos. Na data que este trabalho foi realizado, o DVS de menor custo tinha um valor de mercado de R$8.500,00. Para resolver um problema, é importante que a solução proposta não seja apenas plausível mas também economicamente viável, permitindo que pesquisadores e desenvolvedores com diferentes graus de investimentos consigam usufruir da abordagem em quesão.

Por mais que existam diversas abordagens para tentar resolver o problema de classificação de movimentos, a maioria está relacionada a movimentos do corpo humano. Alguns trabalhos como o de Shaw, Elias e Velusamy (2021) se propõem a solucionar a classificação de movimento de objetos, mas nenhum deles consegue resolver esta tarefa para qualquer tipo de movimento, seja realizado por uma pessoa, animal ou objeto. Isso acontece porque modelos baseados em extração de características até então não eram robustos o suficiente para manter as características espaciais de imagens. Mesmo utilizando redes pré-treinadas para extração de características, a posição de objetos no contexto da imagem era perdida. A forma de resolver isso é através dos chamados modelos fundacionais ou modelos base.

Recentemente alguns trabalhos surgiram que trazem a proposta de serem modelos fundacionais da inteligência artificial. Para a área de visão computacional, destaca-se o artigo proposto por Oquab et al. (2023). Eles propuseram a rede DinoV2, um modelo baseado em transformadores que foi treinado com 1,2 bilhão de imagens únicas, clusterizadas nas classes do conjunto de dados ImageNet através de uma rotina auto-supervisionada. Neste trabalho eles provaram a eficiência de se utilizar modelos auto-supervisionados para tarefa de classificação de imagens. O modelo mais completo possui 1,1 bilhão de parâmetros e serve como base para realização de diversas tarefas. Os autores disponibilizaram quatro modelos diferentes, o que chamam de ”backbones”, baseados nas arquiteturas dos ”vision transformers”ViT-S/14, ViT-B/14, ViT-L/14 e ViT-g/14. A diferença entre eles é o vetor de características produzido na saída das redes. A tabela 2.1 mostra o número de elementos dos vetores de características para cada versão dos modelos disponibilizados.

Versão do modelo	Número de elementos
ViT-S/14	384
ViT-B/14	768
ViT-L/14	1024
ViT-g/14	1536

Tabela 2.1: Número de elementos do vetor de características gerado pelas redes DinoV2.

A escolha do modelo depende do compromisso entre processamento computacional e qualidade do resultado. A partir da escolha do modelo, pode-se acoplar novas camadas

na rede para realizar tarefas como segmentação semântica, estimação de profundidade, recuperação de instâncias, correspondência densa e esparsa. Como o modelo da rede DinoV2 é um excelente extrator de características, ele foi escolhido para o propósito deste trabalho, uma vez que, devido à forma que foi criado, os vetores de características consideram as questões espaciais das imagens. Utilizando a rede DinoV2 como extrator de características em conjunto com uma rede convolucional com memória e uma rede densa, é possível classificar quaisquer tipos de movimentos. Desta forma, esta abordagem se faz proeminente para resolver a tarefa de classificação de movimentos.

A rede DinoV2 é a mais nova versao de sua precursora, a Dino, também criada pela equipe de pesquisas da empresa Meta. Conforme o estudo de Meinardus (2023), o ”pipeline”usado para treinar a rede Dino faz parte da família de técnicas de ”knowledge distillation”chamada ”self-distillation”criada por Zhang, Bao e Ma (2022). De forma resumida, cada imagem entra neste ”pipeline”e passa por dois diferentes processos de transformação aleatória. Os resultados de cada conjunto passam por duas redes ligeiramente diferentes, uma chamada professor e outra aluno (também vistas na literatura com o nome de redes ”target”e ”online”, respectivamente). A rede professor recebe imagens contendo ao menos 50% do conteúdo da imagem original, as chamadas ”global crops”(lembrando que ela poderá ter sido girada, invertida ou mesmo ter suas cores alteradas no processo de transformação aleatória). Já a rede aluno recebe as ”global crops”e também recebe as chamadas ”local crops”, que são imagens que possuem 50% ou menos do conteúdo da imagem original. A rede professor recebe a média móvel exponencial da rede aluno, além de processar as ”global crops”. O que faz esse ”pipeline”especial é o fato da rede professor dividir as imagens recebidas em imagens menores e verificar dentro de cada imagem menor se existe algum objeto interessante para detectar. Isso previne que apenas um objeto seja detectado por imagem, mantendo os aspectos espaciais das mesmas. As saídas das redes professor e aluno acontecem após camadas de softmax, então os vetores resultantes são comparados através de uma função de custo entropia cruzada, função esta que serve como diretriz para o treinamento da rede. Finalizada a técnica de ”self-distillation”, os resultados são colocados em uma arquitetura de ”vision transformer”, ou ViT, e então inicia-se o treinamento auto-supervisionado. A rede DinoV2 acompanha este ”pipeline”, e traz melhorias em alguns aspectos, como no pré-processamento de dados onde é usada uma sofisticada técnica para curadoria dos dados.

Capítulo 3

Métodos

Para atingir os objetivos que este trabalho propõe, os videos com os movimentos que se deseja classificar são processados a fim de serem decompostos em vetores de características, e estes vetores são utilizados como entrada em um modelo de inteligência artificial. Tal modelo faz a predição do movimento, uma vez que seus parâmetros foram treinados para esta tarefa. A seguir são descritos os processos relacionados ao desenvolvimento desta solução, visitando os principais aspectos técnicos da abordagem proposta.

3.1 Procedimentos

Inicialmente um conjunto de dados foi criado a partir de videos coletados pelo autor (mais detalhes sobre o conjunto de dados podem ser encontrados na seção 3.2). Tal conjunto de dados é composto por 1.500 amostras, cada uma delas contendo 19.968 elementos, referentes às três classes diferentes de movimentos. Este conjunto de dados foi dividido em três subconjuntos, um para treinamento, outro para validação e finalmente um para testes. Antes da separação, o conjunto de dados foi embaralhado. O conjunto de treinamento ficou com 1050 amostras, enquanto que o de validação e testes ficaram com 225 amostras cada. Antes de prosseguir com o treinamento de um modelo utilizando tais dados, foi analisada a média e desvio padrão, notando-se que os dados precisam ser normalizados antes de prosseguir. Feita a normalização dos dados foi verificado se os conjuntos estavam balanceados. Feitas as devidas análises, foi possível partir para o treinamento do modelo.

O primeiro modelo treinado possui duas camadas convolucionais de uma dimensão (uma vez que os dados a serem processados são vetores numéricos), cada uma seguida de uma camada de normalização de batelada e função de ativação Unidade Linear de Retificação (ReLU). Após estes dois conjuntos de camadas, o resultado é processado por uma camada ”Long Short-Term Memory”(LSTM). Esta camada é usada para manter o aspecto temporal das informações, uma vez que a sequência de entrada é referente às

características dos quadros dos videos e possuem uma ordem temporal. Após esta camada é utilizada uma camada de ”average pooling”, que obtém as médias dos valores recebidos da camada anterior. Finalmente é utilizada uma camada densa com três unidades e ativação ”softmax”para calcular as probabilidades dos possíveis movimentos. Infelizmente este modelo não obteve bons resultados, os valores de custo e exatidão ficaram instáveis durante o treinamento.

O segundo modelo utilizado possui dois blocos com camadas convolucionais, função de ativação LeakyReLU e camada de ”dropout”. Após as essas duas camadas, o resultado é processado por uma camada de atenção e normalização por batelada, seguida novamente de uma camada de ativação LeakyReLU e ”dropout”. O resultado então é processado por uma camada densa com 50 unidades, seguida de ”global average pooling”e mais uma camada densa com 3 unidades para predição do movimento em questão. Por sua complexidade, em pouquíssimas épocas este modelo atingiu ”overfitting”. Sendo assim, os parâmetros da rede acabaram não sendo efetivamente treinados para conseguir fazer boas predições. Após a investigação deste modelo, foi possível perceber que uma rede mais complexa talvez não seria a melhor abordagem para resolver a tarefa em questão. Assim chegou-se à terceira abordagem de modelo.

O terceiro modelo utilizado possui apenas camadas densas. São utilizadas sete camadas densas, com a quantidade de unidades variando de 32 a 256. Antes da camada final com três neurônios, uma camada de ”global average pooling”é usada. Esta rede apresentou um treinamento mais estável que as demais redes treinadas até então, diminuindo de forma quase constante o custo, ao mesmo tempo que a exatidão aumenta para os conjuntos de treinamento e testes. Após pouco mais de 600 épocas de treinamento, os resultados da rede estagnaram sem melhorias significativas. Por se tratar de uma rede puramente densa, o fato de não haver camadas convolucionais prejudicou o resultado, uma vez que as camadas convolucionais são boas extratoras de características. Um quarto modelo de rede foi criado, unindo camadas densas com camadas convolucionais, bem como camadas de ”dropout”. Mais uma vez a rede rapidamente atinge ”overfitting”. Estes testes com diversos tipos de camadas de rede, bem como diferentes números de unidades nas camadas densas ou mesmo diferentes números de filtros para as camadas convolucionais foram fundamentais para se obter uma rede com melhor resultado para a tarefa proposta. Com tais testes foi possível chegar ao quinto modelo de rede testado.

O quinto modelo criado utiliza camadas convolucionais de uma dimensão seguidas da função de ativação ReLU. Neste caso não é utilizada a camada de normalização de batelada, pois esta camada estava provocando instabilidade no treinamento. É utilizada uma camada recorrente LSTM seguida de ”global average pooling”. Em sua saída é usada uma camada densa com três unidades e função de ativação ”softmax”para predição dos

resultados. Esta rede com pouco mais de 50.000 parâmetros foi capaz de atingir bons resultados com os dados utilizados.

3.2 Materiais

Visando ter maior controle sobre os dados, principalmente no momento da inferência, um conjunto de dados é criado para treinamento, validação e testes. Foram coletados 1.500 videos, cada um com 2 minutos de duração. Cada video é dividido em 26 imagens ordenadas, mantendo-se assim o aspecto temporal dos movimentos. Este número de quadros é escolhido após análise visual dos mesmos, ao observar que os movimentos a serem identificados podem ser vistos com esta quantidade de imagens.

As figuras 3.1 e 3.2 mostram exemplos de videos onde uma mão faz um movimento de ”girar um botão imaginário”. São mostrados os quadros de número 6 a 26 (os primeiros quadros foram suprimidos dessa visualização pois normalmente não contém informações relevantes). Observa-se que a região onde aparece a mão nos quadros representa 22% da imagem original. As figuras 3.3 e 3.4 mostram exemplos de video para o movimento de ”fechar a mão”. Esses movimentos são duas das três classes criadas nos 1.500 videos, sendo que a terceira classe é de videos sem movimentos. Como pode ser visto, entre dois exemplos do mesmo tipo de movimento, a tonalidade da cor muda devido ao horário que o video foi gravado, mas mais do que isso, o momento que os movimentos ocorrem não é fixo. Enquanto na figura 3.1 a mão completa aparece na quinta imagem mostrada, na figura 3.2 somente aparece na nona imagem.

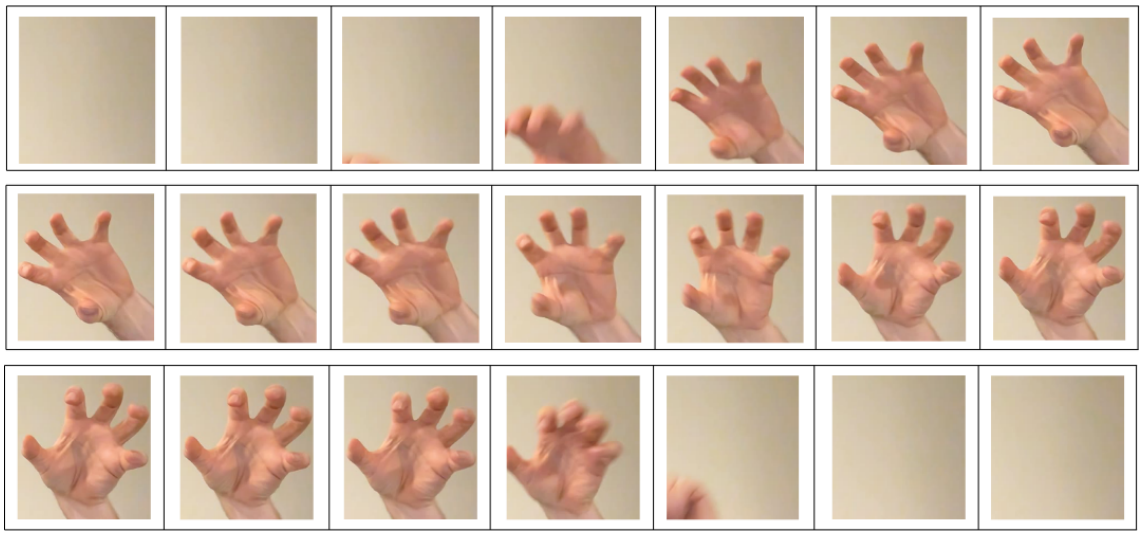

Figura 3.1: Quadros de um exemplo de video captado com o movimento ”girando botão”.

Poderiam ser utilizadas mais imagens para detectar movimentos com maior número de detalhes, porém isso faria com que o vetor final de características ficasse muito grande.

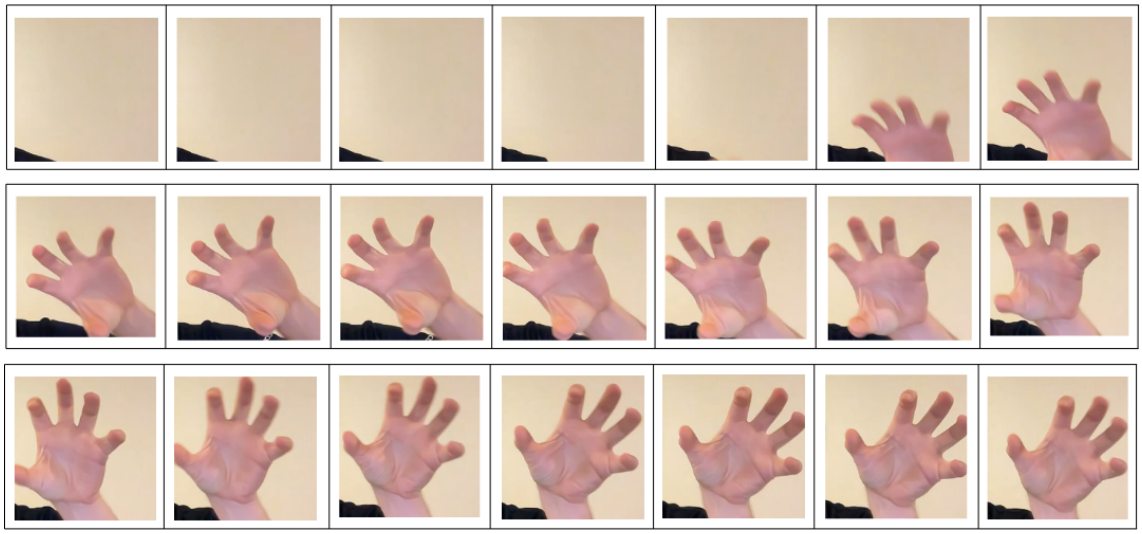

Figura 3.2: Outro exemplo de regiões de um video captado com o movimento ”girando botão”.

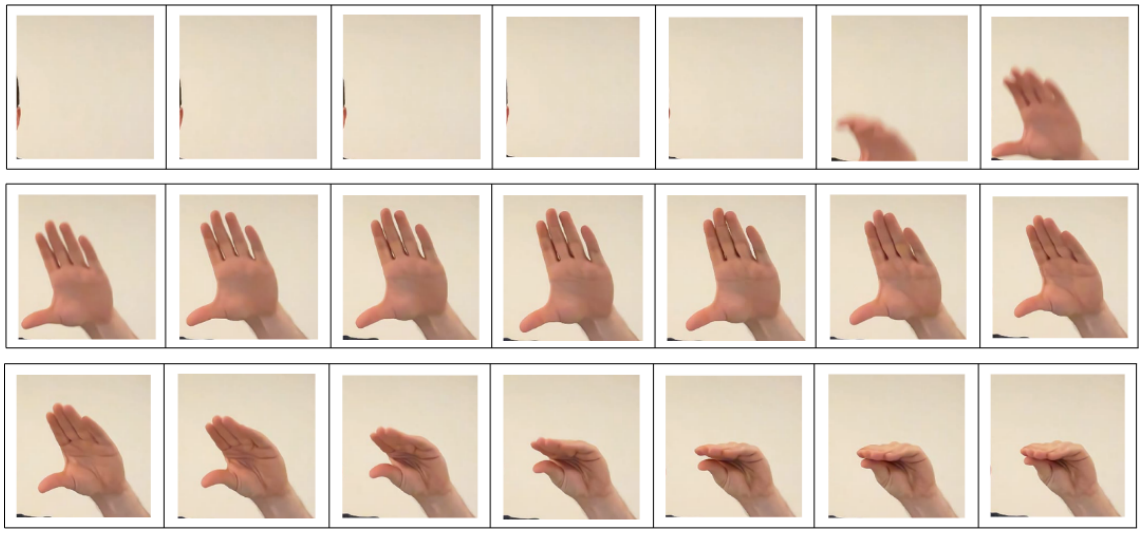

Figura 3.3: Quadros de um exemplo de video captado com o movimento ”fechando a mão”.

Para determinados movimentos, é possível optar por um número menor de imagens, o que faria o processo de treinamento e inferência mais rápidos. Novamente, a escolha da quantidade de quadros depende do compromisso entre a quantidade de detalhes que o movimento possui e a capacidade computacional necessária para a execução da tarefa.

Para compor um conjunto de dados com um bom número de características, foi escolhido o modelo ViT-B/14 da DinoV2, que gera um vetor de características com 768 elementos para cada imagem. Considerando que cada video é composto por 26 imagens, e cada imagem gera um vetor de 768 elementos, cada exemplo do conjunto de dados possui um total de 19.968 elementos. Para automatizar a criação do arquivo CSV do conjunto de dados é criado um algoritmo que grava os videos de 2 minutos após ser apresentada a palma da mão aberta para a câmera; os videos são cortados em quadros ordenados; cada

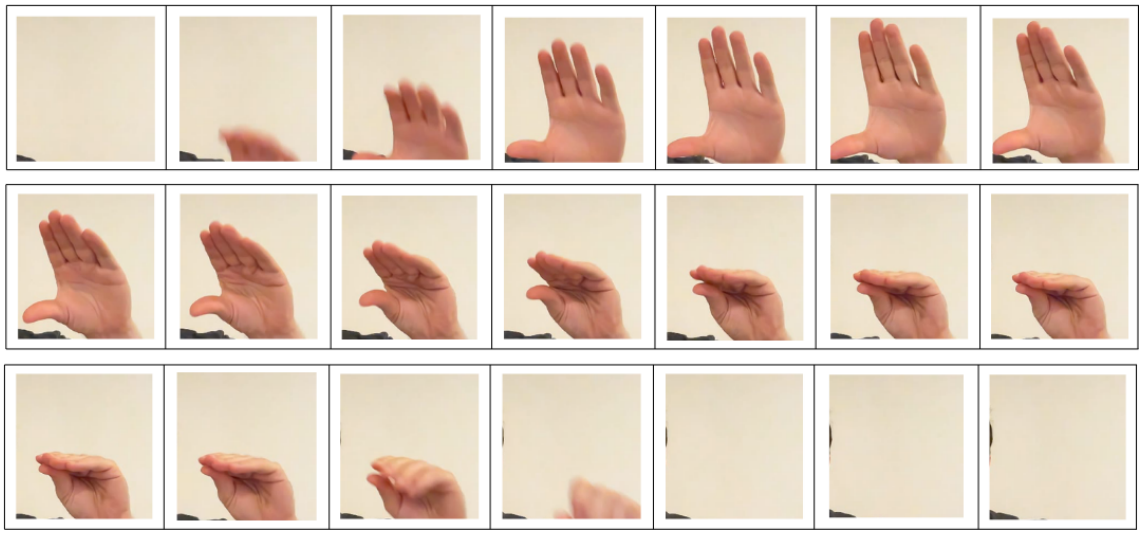

Figura 3.4: Outro exemplo de regiões de um video captado com o movimento ”fechando a mão”.

quadro passa pelo extrator de características ViT-B/14; é gerada uma lista ordenada com todos os vetores de características e finalmente esta lista é alinhada em um vetor de uma linha. Tal vetor é inserido em um arquivo CSV para que posteriormente seja usado para treinamento, validação ou teste da rede.

A figura 3.5 exemplifica de uma forma lúdida o processo utilizado para criação dos dados. Já a figura 3.6 exemplifica o caminho que os dados da série temporal percorre até as unidades finais da rede neural, onde é calculada a probabilidade do video mostrar um determinado movimento.

3.3 Instrumentos

Para coleta dos videos, criação do conjunto de dados e testes de inferência foi utilizado um computador Apple Macbook com processador M1 Pro, contendo 16GB de memória unificada, compartilhada entre RAM e GPU. Para o treinamento do modelo são usadas as máquinas disponíveis no Google Colab. Como diversos modelos foram criados durante os testes, ora foi utilizada uma placa de vídeo A100, ora foi utilizada uma placa V100.

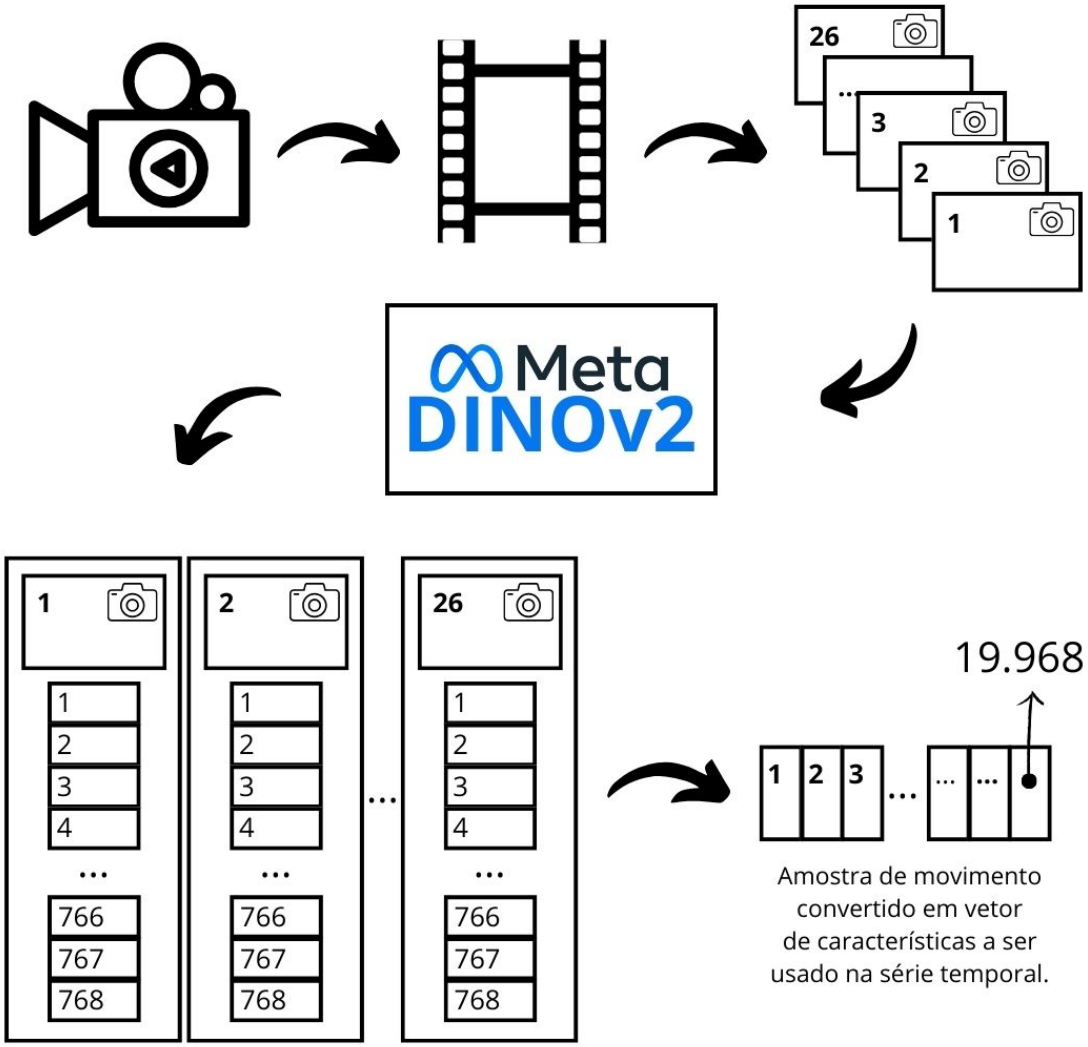

Figura 3.5: Processo de captura de videos e criação do vetor de características, utilizando a rede DinoV2.

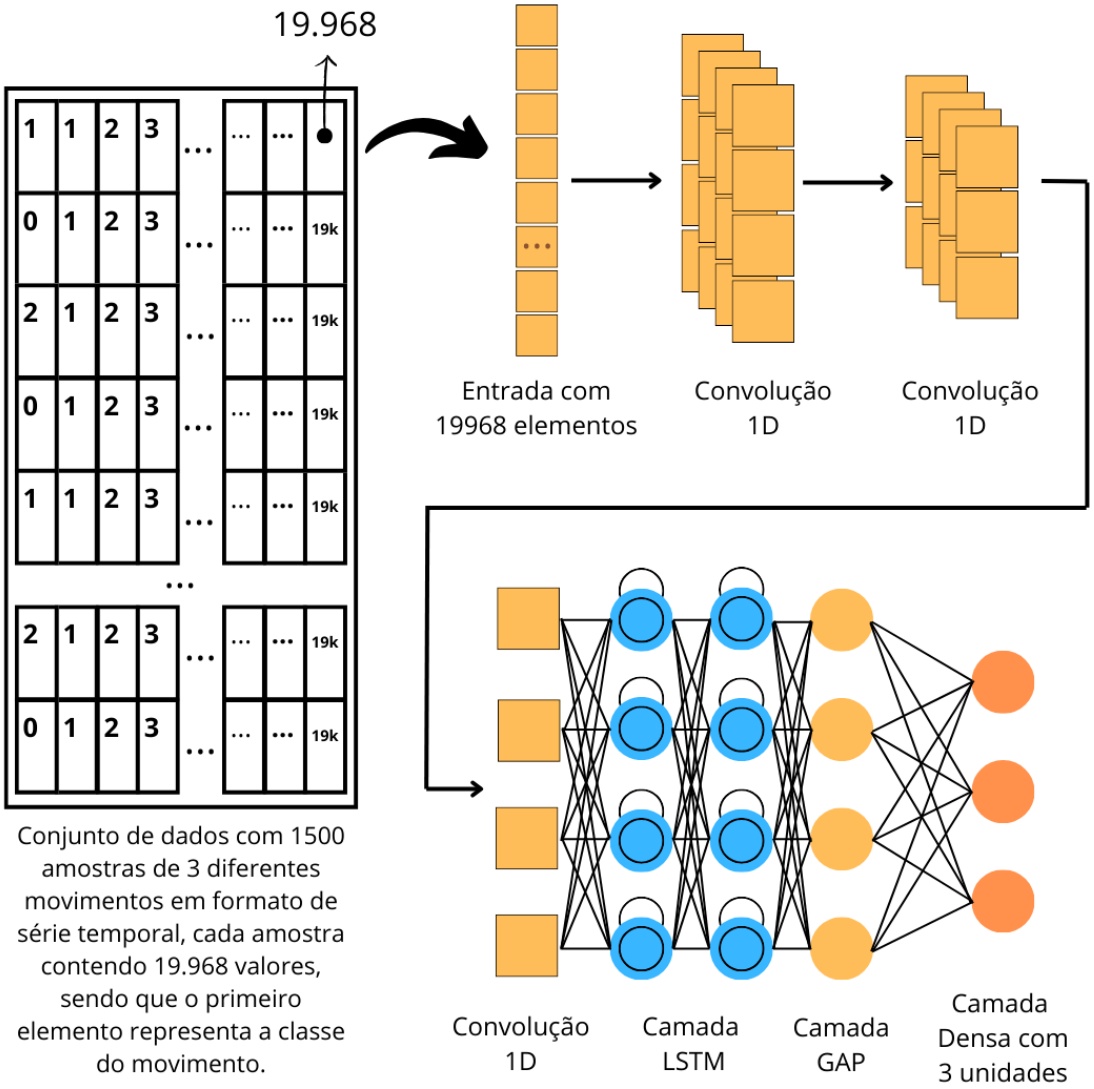

Figura 3.6: Caminho que as amostras do conjunto de dados percorrem para classificação dos movimentos, utilizando camadas convolucionais, LSTM e densas.

Capítulo 4

Resultados e Discussão

Por mais que o modelo não seja treinado com um conjunto de dados exaustivo, os resultados preliminares são promissores. A exatidão no conjunto de treinamento foi de 94,3%, enquanto que no conjunto de teste foi de 93,3%. O valor da função de custo no treinamento ficou em 0,293 e em 0,319 para o conjunto de testes. Mesmo com um custo relativamente elevado, o modelo resultante apresentou bons resultados durante a inferência de dados nunca vistos anteriormente. A matriz de confusão mostrada na Figura 4.1 mostra os resultados para o conjunto de testes.

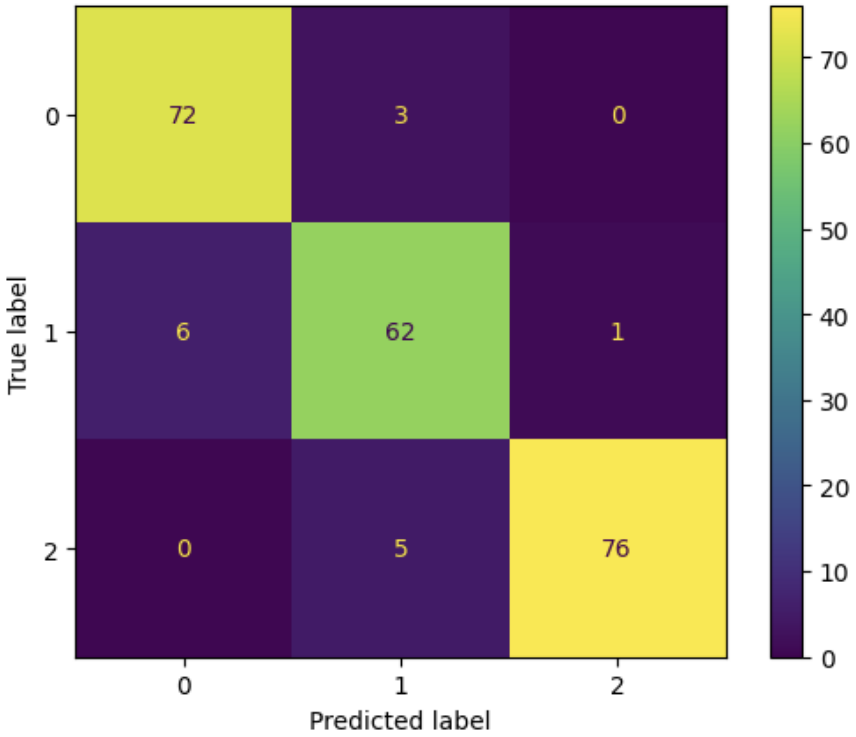

Figura 4.1: Matriz de confusão do último modelo criado.

Considerando que o conjunto de dados foi criado com uma única pessoa, e que esta pessoa estava sempre na mesma posição com pequenas diferenças na cor de seu vestuário, o modelo faz a previsão correta na maioria das inferências com dados nunca antes vistos. Foram feitos mais de 100 testes na mesma posição que o conjunto de dados original foi gerado, e o modelo acertou o movimento em mais de 95% dos casos.

Os gráficos da figura 4.2 mostram as curvas de custo e exatidão do treinamento do modelo, ao longo das 56 épocas que foi treinado. Devido a uma parada prematura tendo como referência o valor da exatidão para os dados de validação, o modelo encerrou seu treinamento após este valor ser superior a 95%. Devido ao fato de não ter sido treinado durante muitas épocas, mesmo com tamanho do lote de 32 amostras, os valores dos custos de treinamento e validação ficaram elevados. Sendo assim, o resultado poderia ter sido melhor caso alguns parâmetros da rede fossem alterados, e se tivesse sido treinada por mais épocas.

Utilizando modelos com diferentes configurações, conforme descrições apresentadas no capítulo anterior, os resultados obtidos foram piores. Nota-se uma clara necessidade do uso de ao menos uma camada LSTM para melhoria dos resultados. Isso era esperado, já que esta arquitetura de rede neural recorrente possui mecanismos de memória que são muito bem-vindos para classificar, processar e prever séries temporais. Em uma camada recorrente simples, durante a propagação para trás valores muito distantes do momento atual tendem a zerar, pelo efeito de ”vanishing gradients”. As camadas LSTM possui mecanismos que ”entendem”o que é ou não importante durante o processo de treinamento, e mantém apenas informações importantes. O que determina se as informações são importantes são os chamados portões, que nada mais são que camadas densas internas à camada LSTM. Modelos sem este tipo de camada perdem informações durante o treinamento, fazendo com que seus resultados não sejam relevantes.

Outro escolha importante para se obter bons resultados foi a do modelo DinoV2 para extração de características. Devido à arquitetura e forma de treinamento da DinoV2, ela se torna extremamente poderosa para a criação dos vetores de características de cada imagem. Redes tradicionais de classificação de imagens como a proposta por Simonyan e Zisserman (2014) são normalmente baseadas em redes convolucionais profundas. Seja a VGG16, seja a VGG19, a camada de saída dessas redes não preserva os aspectos espaciais das imagens de entrada. O modelo DinoV2 consegue preservar tais informações, e as mesmas podem ser usadas para segmentação de imagens e outras operações como a proposta neste documento.

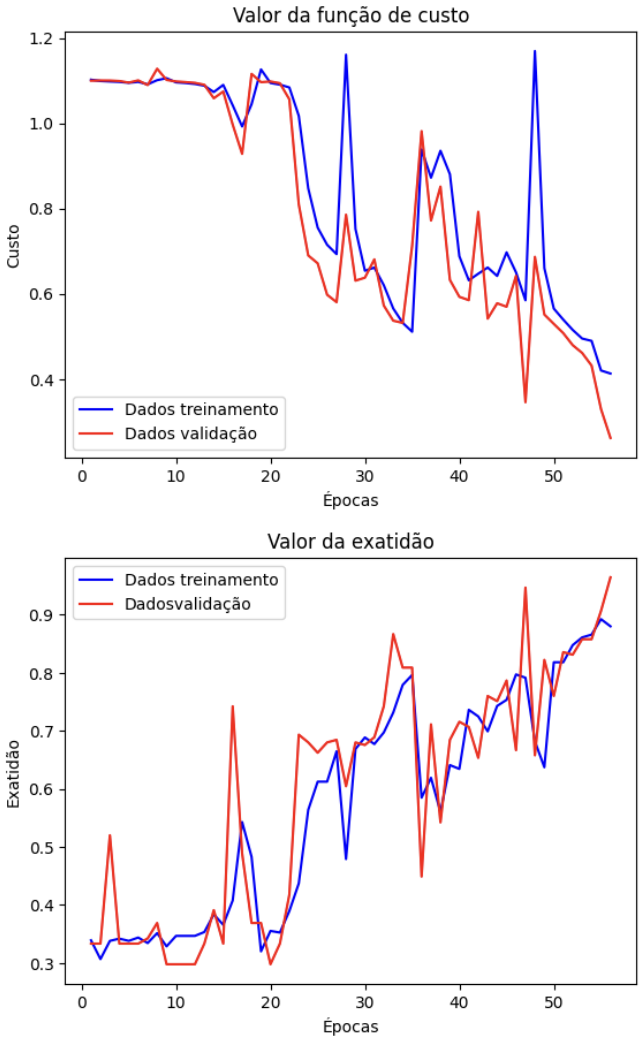

Figura 4.2: Custo e exatidão do modelo criado.

Capítulo 5

Conclusão

O objetivo desse trabalho é encontrar uma forma de classificar movimentos genéricos em videos. Utilizando um modelo fundacional para extração de características juntamente a uma rede com camadas convolucionais, recorrentes e densas, é possível iniciar um estudo promissor sobre o tema. Devido à limitação do conjunto de dados utilizado, é cedo para dizer que esta forma de classificação resolve a tarefa proposta, mas os resultados são animadores.

Com a realização desse estudo, novos trabalhos poderão ser realizados com o intuito de dar continuidade a este desenvolvimento. Os próximos passos para este projeto seriam o treinamento de novos modelos com outros tipos de movimento, ou mesmo o uso de conjuntos de dados que tornem o modelo atual mais genérico. As camadas de classificação ainda podem ser melhoradas, encontrando melhores parâmetros ou mesmo adicionando outras camadas propícias a séries temporais.

Assim como ocorreu com os modelos de linguagem, os modelos fundacionais voltados à imagens devem ser vistos como uma poderosa ferramenta para os mais diversos tipos de desenvolvimento no campo da visão computacional. Tais modelos aceleram a criação de novas soluções, como ocorreu recentemente com o BERT e o GPT-n para linguagem. Provavelmente em breve serão desenvolvidas ideias revolucionárias no campo da visão computacional, advindas da junção das diversas técnicas de inteligência artificial com os modelos fundacionais.

Referências

Ahmed, Wafaa Shihab et al. (2020). "Motion classification using CNN based on image difference". Em: *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*. IEEE, pp. 1–6.

Amir, Arnon et al. (2017). "A low power, fully event-based gesture recognition system". Em: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7243–7252.

Meinardus, Boris (ago. de 2023). *Self-Supervised Learning and Transformers? — DINO Paper Explained*. URL: https://pub.towardsai.net/self-supervised-learning-and-transformers-dino-paper-explained-8fd7458ce2c9.

Oquab, Maxime et al. (2023). "Dinov2: Learning robust visual features without supervision". Em: *arXiv preprint arXiv:2304.07193*.

Shaw, Soumya, Susan Elias e Sudha Velusamy (2021). "Feature Engineering for Motion Classification in Machine Vision". Em: *Journal of Physics: Conference Series*. Vol. 2115. 1. IOP Publishing, p. 012043.

Simonyan, Karen e Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". Em: *arXiv preprint arXiv:1409.1556*.

Xing, Yannan, Gaetano Di Caterina e John Soraghan (2020). "A new spiking convolutional recurrent neural network (SCRNN) with applications to event-based hand gesture recognition". Em: *Frontiers in neuroscience* 14, p. 590164.

Zhang, L., C. Bao e K. Ma (2022). "Self-Distillation: Towards Efficient and Compact Neural Networks". Em: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.