UNIVERSIDADE FEDERAL DE OURO PRETO
DEPARTAMENTO DE CIÊNCIA DE COMPUTAÇÃO
PROCESSAMENTO DIGITAL DE IMAGENS

Projeto de Pesquisa

SKELETON BASED ACTION RECOGNTION PARA RECONHECIMENTO DE LINGUAGEM DE SINAIS

Alunos:

Beatriz Helena de Mello Orlandi de Deus,

Caio Silas de Araujo Amaro,

Julia Carlos Gonzaga,

Josué Villa Real

Resumo

Este projeto investiga uma abordagem de reconhecimento de sinais em Língua Brasileira de Sinais (LIBRAS) utilizando dados de esqueleto. Onde, propõe-se o uso de informações tridimensionais da base LibrasUFOP, extraídas por sensores (como o Microsoft Kinect) e algoritmos modernos de estimativa de pose para capturar de forma robusta os gestos. O trabalho enfatiza a integração de técnicas de pré-processamento, arquitetura Transformer com Multi-Head Attention, a qual possibilita o processamento paralelo e a captura de relações temporais de longo alcance. O sistema proposto é eficiente e de baixo custo computacional, com o objetivo de aprimorar a comunicação e promover a inclusão social da comunidade surda.

1 Introdução

A linguagem de sinais é um meio essencial de comunicação para pessoas com deficiências auditiva e alguns tipos de deficiências físicas, como nas cordas vocais, permitindo a interação tanto entre deficientes auditivos quanto com ouvintes em diversas situações do cotidiano. Atualmente, grande parte dos serviços de tradução dessa linguagem ainda depende de intérpretes humanos, o que pode ser oneroso devido à necessidade de profissionais altamente qualificados.

O reconhecimento automático da linguagem de sinais apresenta desafios únicos, que vão além do simples reconhecimento de ações. Ele exige a identificação precisa de movimentos corporais e expressões faciais, pois gestos semelhantes podem ter significados distintos dependendo do contexto. Além disso, cada indivíduo possui variações próprias de velocidade, estilo e regionalismos, tornando o processo ainda mais complexo (Jiang et al., 2021). A coleta de um grande volume de dados de diferentes signatários, embora essencial, é trabalhosa e dispendiosa.

As dificuldades associadas ao aprendizado e à interpretação da língua de sinais refletem uma problemática maior: as barreiras de comunicação enfrentadas pela comunidade surda. A escassez de intérpretes em instituições públicas e privadas compromete o acesso a direitos fundamentais, como saúde e educação, afetando diretamente a qualidade de vida dessas pessoas (Souza et al., 2017). Esse cenário reforça a necessidade de soluções tecnológicas que promovam maior inclusão e acessibilidade.

Nos últimos anos, o reconhecimento de gestos e atividades humanas tem ganhado crescente interesse da comunidade científica, especialmente devido ao potencial de aplicação na tradução automática da língua de sinais (De Souza et al., 2010). Técnicas avançadas de visão computacional e aprendizado de máquina vêm sendo empregadas para aprimorar a interpretação dos gestos e integrar pessoas surdas à sociedade de maneira mais eficiente.

A inteligência artificial tem sido uma grande aliada nesse avanço. A Organização Mundial da Saúde (OMS) publicou um relatório global destacando os progressos e os benefícios da IA na saúde (OPAS/OMS, 2021), evidenciando o impacto positivo dessa tecnologia em diversos setores, incluindo a acessibilidade para deficientes auditivos.

Diversas abordagens têm sido exploradas para melhorar a acurácia no reconhecimento da linguagem de sinais. Enquanto métodos baseados em imagens podem ser computacionalmente exi-

gentes e menos robustos diante de variações de iluminação e fundo, a utilização de dados do esqueleto humano tem se mostrado promissora. Sensores como o Microsoft Kinect (Zhang, 2012) e algoritmos avançados de estimativa de pose humana facilitam a captura de dados tridimensionais (3D), permitindo um reconhecimento mais preciso dos gestos.

O aprimoramento contínuo dessas tecnologias abre caminho para sistemas mais eficientes e acessíveis, reduzindo as barreiras de comunicação enfrentadas pela comunidade surda e promovendo uma sociedade mais inclusiva.

2 Objetivos

2.1 Objetivo Geral

Desenvolver um sistema de reconhecimento de linguagem de sinais baseado na análise de dados de esqueleto, capaz de distinguir com alta acurácia os sinais da Língua Brasileira de Sinais (LIBRAS) por meio do processamento de séries temporais de coordenadas corporais utilizando uma arquitetura Transformer.

2.2 Objetivos Específicos

- **Coleta e Pré-processamento:** Implementar um método robusto para a aquisição e limpeza de sequências de esqueleto, garantindo a qualidade dos dados capturados.
- **Extração de Características:** Converter os dados brutos em vetores temporais, enriquecidos com informações de velocidade e aceleração (derivadas temporais), para realçar os movimentos discriminativos.
- **Desenvolvimento do Modelo:** Projetar e treinar um modelo Transformer com mecanismo de Multi-Head Attention, de forma a capturar relações temporais complexas e preservar a ordem dos frames por meio de embeddings posicionais.
- **Validação e Teste:** Realizar experimentos com divisão subject-wise (por sujeito), utilizando validação cruzada K-Fold e métricas padrão (acurácia, precisão, recall e F1-score) para avaliar o desempenho do sistema.

- **Análise Comparativa:** Comparar os resultados obtidos com abordagens anteriores e discutir as diferenças metodológicas e de performance.
- **Visualização:** Desenvolver ferramentas de visualização, como curvas de perda e matrizes de confusão, para facilitar a interpretação dos resultados e a identificação de erros.

3 Estado da Arte

O reconhecimento de ações humanas baseado em esqueleto tem sido amplamente explorado nos últimos anos, particularmente devido à sua relevância em aplicações como o reconhecimento de gestos e de linguagem de sinais. Diferentemente de métodos baseados em vídeo RGB, os dados de esqueleto fornecem uma representação topológica detalhada do corpo humano, composta por articulações e ossos, sendo menos sensíveis a variações de aparência e condições ambientais. Nesta seção, apresentamos uma análise crítica de alguns trabalhos relevantes que avançam o estado da arte no reconhecimento de ações humanas baseado em esqueleto.

3.1 Trabalhos Relacionados

Jiang et al. Jiang et al. (2021) propõem o Skeleton Aware Multi-modal SLR (SAM-SLR), que combina esqueleto, RGB e profundidade (RGB-D) para melhorar a acurácia no reconhecimento de linguagem de sinais. O modelo integra a Sign Language Graph Convolution Network (SL-GCN), que modela dinâmicas esqueléticas, e a Separable Spatial-Temporal Convolution Network (SSTCN), que aprimora a extração de características, atingindo 98,42% de acurácia em RGB e 98,53% em RGB-D no AUTSL. Sua estratégia de redução de grafos melhora a eficiência computacional e supera métodos anteriores baseados apenas em ST-GCN, enquanto a fusão multimodal das informações de esqueleto, fluxo óptico e profundidade aumenta a robustez. No entanto, a dependência de grandes volumes de dados e a qualidade da estimativa de poses ainda são desafios, indicando que futuras pesquisas devem explorar arquiteturas mais eficientes e aprendizado semi-supervisionado para reduzir a necessidade de dados anotados.

Por outro lado, Ye et al. (Ye et al., 2020) introduzem o *Dynamic GCN*, que utiliza grafos dinâmicos enriquecidos por contexto para modelar a topologia do esqueleto humano. O modelo,

por meio do módulo *Context-encoding Network* (CeN), adapta automaticamente as conexões do grafo, considerando interdependências contextuais. Diferentemente de grafos estáticos, o *Dynamic GCN* constrói topologias específicas para cada amostra, combinando-as com topologias estáticas baseadas em conexões físicas. Essa abordagem alcançou 91,5% de acurácia no NTU60 (*Cross-Subject*) e 96,0% (*Cross-View*), com um aumento de apenas ~7% nos FLOPs em relação ao modelo base. Contudo, a geração de grafos dinâmicos adiciona complexidade ao treinamento e maior sensibilidade a ruídos nos dados.

Mais recentemente, Duan et al. (2022) propuseram o PoseConv3D, uma abordagem inovadora que revoluciona o campo ao utilizar volumes de heatmap 3D em substituição às tradicionais sequências de grafos. O método se destaca por sua capacidade de processar poses 2D através de estimadores modernos e convertê-las em mapas de calor empilhados temporalmente, oferecendo maior robustez contra ruídos na estimação de poses. Uma das principais contribuições do PoseConv3D é sua capacidade de processar múltiplas pessoas simultaneamente sem custo computacional adicional, além de permitir uma integração mais natural com outras modalidades. A arquitetura demonstrou superioridade ao alcançar estado da arte em cinco de seis benchmarks padrão, estabelecendo novos patamares de performance no reconhecimento de ações baseado em esqueleto. Sua eficácia é particularmente notável quando combinada com outras modalidades, tendo obtido os melhores resultados em todos os oito benchmarks de reconhecimento multimodal avaliados.

Hu et al. Hu et al. (2024) propõem o Dynamic Spatial-Temporal Aggregation (DSTA-SLR), um método para reconhecimento de linguagem de sinais baseado em esqueleto que supera limitações de abordagens anteriores ao modelar relações dinâmicas entre articulações e capturar padrões temporais complexos. O modelo utiliza um módulo de correlação de grafos, que ajusta conexões esqueléticas de forma sensível à entrada, e um módulo de convolução temporal paralela, que extrai informações multi-escalas de movimentos. Nos benchmarks WLASL, MSASL, SLR500 e NMFs-CSL, a abordagem alcançou estado da arte, superando métodos baseados em esqueleto e RGB na maioria dos casos, enquanto mantém menor demanda computacional. A modelagem dinâmica das conexões articulares e a fusão de múltiplos fluxos de informação aumentam a robustez do modelo, mas a qualidade da estimativa das poses e a necessidade de dados extensivos ainda são desafios. Trabalhos futuros podem explorar estratégias de aprendizado mais eficientes para reduzir a dependência de anotações manuais.

O estudo de Özdemir et al. (2023) apresenta o modelo Multi-Cue Temporal Modeling (MCTM), uma abordagem inovadora para o reconhecimento de linguagem de sinais com base em sequências de esqueleto. O MCTM se destaca por integrar diferentes pistas temporais, permitindo capturar dependências complexas nos movimentos dos sinais, ao mesmo tempo em que combina informações espaciais e temporais. Essa abordagem visa melhorar a precisão do reconhecimento, ao tratar as interações dinâmicas dos gestos com mais profundidade. A arquitetura proposta combina redes neurais recorrentes e convolucionais, com o intuito de modelar tanto as dinâmicas temporais quanto as espaciais presentes nas sequências de esqueleto. Nos experimentos realizados, o MCTM superou significativamente os métodos anteriores, alcançando uma impressionante acurácia de 92,3% no dataset CSL-Daily e 84,7% no dataset SLR500. No entanto, apesar de seu desempenho superior, o modelo exige uma quantidade substancial de dados de treinamento para aprender as variações sutis nos gestos da linguagem de sinais, o que pode limitar sua aplicação em cenários com recursos de dados mais escassos.

Recentemente, Chen et al. (2024) introduziram o modelo SignVTCL, para o reconhecimento contínuo de linguagem de sinais, fundamentada no conceito de aprendizado contrastivo visual-textual. A estratégia busca otimizar o alinhamento entre representações visuais e textuais, maximizando a similaridade entre pares correspondentes e minimizando a similaridade entre pares não relacionados. Para tal, o modelo emprega uma arquitetura baseada em redes neurais convolucionais para extrair características visuais robustas de múltiplas modalidades, dentre as quais se destacam vídeos RGB, pontos-chave corporais (representando posições articulares) e fluxo óptico, que capta com precisão as dinâmicas temporais entre quadros consecutivos. Ademais, o SignVTCL realiza o alinhamento em dois níveis: o gloss-level, que associa segmentos visuais específicos a glosses individuais, e o sentence-level, que considera o contexto global da sentença, permitindo uma interpretação semântica mais refinada dos sinais contínuos. Nos benchmarks Phoenix-2014, Phoenix-2014T e CSL-Daily, o modelo estabeleceu novos patamares de robustez e generalização, superando abordagens anteriores em cenários marcados por variações estilísticas acentuadas e escassez de dados anotados. Entretanto, a dependência da qualidade das anotações e a complexidade na extração precisa dos pontos-chave ainda se configuram como desafios, sugerindo que investigações futuras devem explorar estratégias de aprimoramento do alinhamento multimodal e métodos de aprendizado semi-supervisionado para mitigar tais limitações.

Cerna et al. Cerna et al. (2021) apresentam o LIBRAS-UFOP, um conjunto de dados multimodal para reconhecimento de sinais em LIBRAS, capturado com o Microsoft Kinect V1 e contendo informações completas de RGB-D e esqueleto. O dataset é baseado no conceito de pares mínimos, agrupando 56 sinais organizados em quatro categorias, validados por um especialista. Para avaliação, os autores propõem um método baseado na geração de imagens dinâmicas a partir dos dados multimodais, utilizadas como entrada para redes convolucionais (CNNs). O modelo alcançou 74,25% de acurácia, demonstrando a complexidade do dataset e superando abordagens tradicionais baseadas em descritores manuais. A fusão de múltiplas modalidades melhora a robustez, mas a similaridade entre sinais ainda representa um desafio. Pesquisas futuras podem explorar redes mais eficientes e estratégias de aprendizado semi-supervisionado para reduzir a dependência de grandes volumes de dados anotados.

Em continuidade ao uso do dataset LIBRAS-UFOP, Alves et al. Alves et al. (2024) propõem uma abordagem inovadora para reconhecimento de sinais isolados em Libras utilizando representação de imagens baseadas em esqueletos. Diferentemente de métodos tradicionais que dependem de dados multimodais complexos e redes convolucionais 3D, o método proposto extrai landmarks corporais, faciais e das mãos utilizando o OpenPose, codificando essas informações espaço-temporais em uma única imagem 2D, que posteriormente é classificada por uma CNN 2D simplificada. Essa abordagem atingiu acurácias superiores aos métodos anteriores nos datasets MINDS-Libras (93%) e LIBRAS-UFOP (82%), destacando-se pela simplicidade da arquitetura e eficiência no treinamento, utilizando exclusivamente dados RGB. No entanto, a dependência da extração de landmarks pelo OpenPose ainda representa uma limitação considerável, devido ao alto custo computacional associado, o que pode restringir aplicações em tempo real. Futuros estudos poderiam explorar técnicas mais rápidas de extração de poses e estratégias para minimizar a perda de precisão associada à aceleração desse processo.

Plaza et al. (2024) apresentam um novo método autorregressivo para a predição de movimento de pedestres, combinando Transformers e Redes Convolucionais de Grafos. O modelo é capaz de analisar a informação temporal e espacial dos esqueletos capturados e prever o movimento futuro. Os experimentos foram realizados sobre o conjunto de dados Kinetics-Skeleton, e os resultados mostraram a eficácia da abordagem na previsão da posição das articulações. A avaliação utilizou métricas como erro quadrático médio (MSE) e erro angular médio (MAE), destacando a impor-

tância da fusão de informações espaciais e temporais. Os autores apontam que futuras pesquisas devem explorar a aplicação do modelo em contextos veiculares e aprimorar sua capacidade de lidar com sequências temporais mais longas, além de considerar a proximidade dos pedestres em relação ao veículo para uma melhor tomada de decisão.

4 Metodologia

Nesta seção, descrevemos detalhadamente a metodologia adotada para o reconhecimento de sinais a partir dos dados de esqueleto, inspirada no trabalho de Cerna et al. (2021), enfatizando a nova implementação baseada em Transformer.

4.1 Aquisição e Pré-processamento de Dados

Coleta de Sequências de Esqueleto e Estrutura da Base de Dados. A base LibrasUFOP é composta por 4 categorias de sinais, cada uma com características específicas:

- **Categoria 1:** Mesmo movimento, mesmo ponto de articulação, mas com diferentes configurações da mão (10 classes).
- **Categoria 2:** Diferente movimento, mesmo ponto de articulação, mesma configuração da mão (19 classes).
- **Categoria 3:** Mesmo movimento, diferente ponto de articulação, mesma configuração da mão (8 classes).
- **Categoria 4:** Sinais que incluem expressões faciais (19 classes).

Nesta implementação, utilizamos apenas os dados de esqueletos das *categorias 1 e 2*. Cada amostra é uma sequência temporal de coordenadas (x, y, z) das principais juntas do corpo humano, extraídas dos arquivos de texto contidos na pasta `Skel` de cada sinal. A estrutura da base de dados organiza os sinais em pastas, onde cada pasta (nomeada segundo o padrão `pX_cY_sZ`) um subdiretório `Skel` com os arquivos de coordenadas 3D para cada *frame*.

Figura 1 mostra três exemplos de sinais diferentes. Os sinais na Figura 1.a e Figura 1.b são da Categoria 1, eles apresentam variações na configuração da mão, no entanto, o movimento e os

pontos de articulação são os mesmos. Os sinais na Figura 1.a e Figura 1.c são da Categoria 2, eles mostram variações no movimento, mas os pontos de articulação e a configuração da mão são os mesmos.

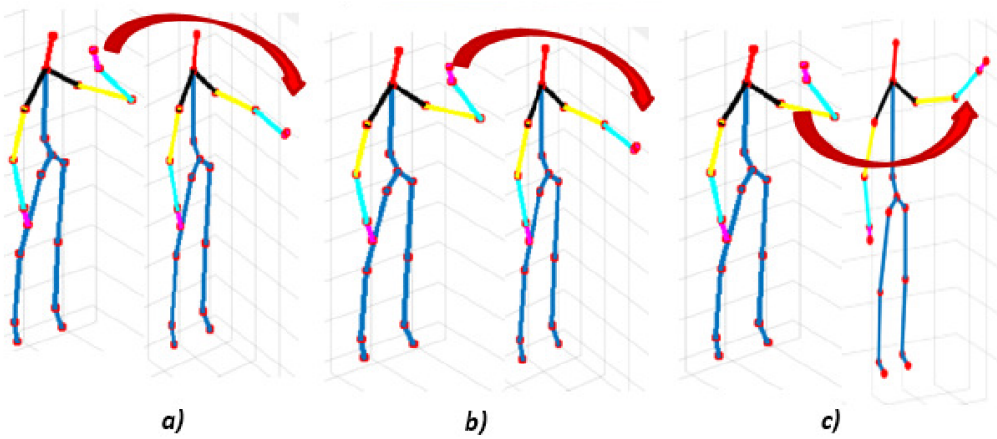

Figura 1: Par mínimo em BSL. Os sinais a) "Dia 1"e b) "Dia 2"representam a Categoria 1. Os sinais a) "Dia 1"e c) "Ideia"representam a Categoria 2. Fonte: Cerna et al. (2021)

Filtragem e Padronização. Inicialmente, os arquivos são lidos por meio da função `load_skeleton_segment`, que extrai as coordenadas de cada junta. Em seguida, a função organiza os frames de acordo com intervalos anotados (start e end) e a função `normalize_skeleton_sequence` centraliza cada frame com base em uma junta de referência (por exemplo, o `SPINE` com índice 2). Essa etapa pode incluir, opcionalmente, uma normalização de escala (dividindo pela distância entre juntas-chave) para reduzir variações intersujeito. A função `fix_joints` garante que cada frame contenha exatamente o número esperado de juntas, preenchendo com zeros ou truncando quando necessário.

Extração de Derivadas. Para capturar informações dinâmicas, a função `compute_temporal_derivatives` calcula a velocidade (primeira derivada) e a aceleração (segunda derivada) entre frames consecutivos. Essa técnica enriquece a representação temporal, permitindo que o modelo detecte mudanças sutis no movimento.

Criação do Conjunto de Dados. A classe `SkeletonSequenceDataset`, derivada de `tf.keras.utils.Sequence`, organiza as amostras em batches para o treinamento. Ela aplica pad-

ding nas sequências (para uniformizar o comprimento) e converte os rótulos em formato *one-hot encoding*. Essa classe possibilita a leitura eficiente dos dados durante o treinamento, suportando operações como embaralhamento e segmentação por sujeito.

4.2 Representação e Extração de Características

Cada *frame* de esqueleto é transformado em um vetor unidimensional, concatenando as coordenadas (x, y, z) das juntas. Com a inclusão das derivadas temporais, o vetor final possui dimensão $3N \times 3$, onde N é o número de juntas. Essa representação rica permite que o modelo capture tanto a posição estática quanto a dinâmica dos movimentos.

4.3 Modelo Transformer para Dados de Esqueleto

Motivação para o Transformer. Embora abordagens baseadas em LSTM com mecanismos de atenção tenham sido exploradas, a arquitetura Transformer oferece vantagens significativas, como o processamento paralelo e a capacidade de capturar relações de longo alcance por meio da *Multi-Head Attention*. Isso é particularmente útil para reconhecer padrões temporais complexos em dados de esqueleto com alta acurácia.

Componentes do Modelo Transformer.

- **Entrada e Máscara:** A camada `Input` define a forma de entrada (T, D), onde T é o comprimento máximo da sequência e D é a dimensão do vetor (incluindo posições, velocidades e acelerações). A camada `Masking` ignora os tokens de padding.
- **Embedding Posicional:** Como os Transformers não incorporam a ordem dos elementos de forma inerente, um *embedding posicional* é adicionado à entrada para preservar a sequência temporal. No código, utiliza-se uma camada `Embedding` aplicada sobre um vetor de posições gerado por `tf.range`. Assim, a entrada é transformada como:

$$\mathbf{X}_{\text{emb}} = \mathbf{X} + \text{PE}$$

onde PE $\in \mathbb{R}^{T \times D}$ é a matriz de embedding posicional, que fornece um vetor único para cada posição da sequência.

- **Multi-Head Attention:** A camada `MultiHeadAttention` realiza o mecanismo de self-attention, permitindo que o modelo se concentre em diferentes subespaços simultaneamente. O mecanismo de atenção é definido por:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

No código, os mesmos dados $\mathbf{X}_{\text{emb}}$ são utilizados para formar as matrizes **Q** (Query), **K** (Key) e **V** (Value), com os parâmetros `num_heads` e `key_dim` controlando o número de cabeças e a dimensão das chaves, respectivamente, veja na figura 2.

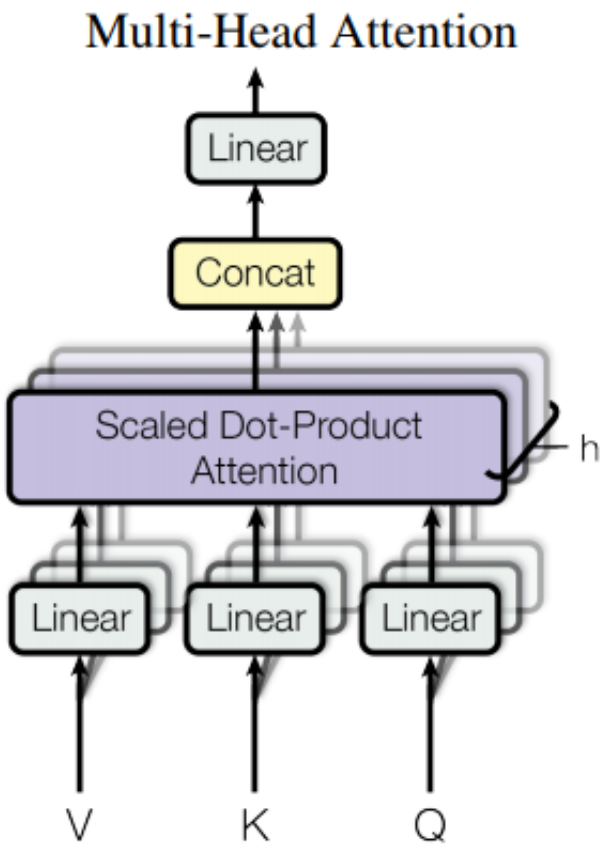

Figura 2: Diagrama da arquitetura de atenção de várias cabeças.

- **Layer Normalization e Conexão Residual:** A saída do mecanismo de atenção é somada à entrada original, formando uma conexão residual que ajuda a estabilizar e acelerar o treinamento. Em seguida, aplica-se a `LayerNormalization`:

$$\mathbf{Y} = \text{LayerNorm}(\mathbf{X}_{\text{emb}} + \text{Attention}(\mathbf{X}_{\text{emb}}, \mathbf{X}_{\text{emb}}, \mathbf{X}_{\text{emb}}))$$

- **Camadas Densas com Regularização:** Após a atenção, o modelo utiliza uma camada densa com ativação `relu` para refinar a representação. São aplicadas técnicas de regularização, como *dropout* e normalização (`BatchNormalization`), além da regularização L2 para prevenir o overfitting. Essa etapa pode ser descrita de forma simplificada como:

$$\mathbf{Z} = \text{BatchNorm}(\text{Dropout}(\text{ReLU}(\text{Dense}(\mathbf{Y}))))$$

- **Global Average Pooling 1D:** Para reduzir a dimensão temporal e gerar uma representação fixa da sequência, é aplicada a operação de *Global Average Pooling 1D*. Essa operação calcula a média dos vetores ao longo dos T frames:

$$\mathbf{z}_{\text{avg}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{z}_t$$

- **Camada Fully Connected Final:** Uma camada densa adicional, seguida de *dropout* e normalização, transforma a representação reduzida em uma forma apropriada para a classificação final.

- **Saída:** A camada final é uma `Dense` com ativação `softmax` e número de neurônios igual ao número de classes, gerando a distribuição de probabilidade para cada sinal.

O modelo é construído pela função `build_transformer_with_attention`, que encapsula esses passos e retorna um objeto `Model` pronto para treinamento.

4.4 Treinamento e Validação Cruzada

Validação Cruzada K-Fold. Para uma avaliação robusta da generalização, foi implementada a validação cruzada utilizando o método K-Fold (com 5 folds). Essa abordagem divide as amostras em diferentes subconjuntos de treino e validação, reduzindo o viés na seleção dos dados.

Callbacks e Otimização. Durante o treinamento, foram empregados callbacks como:

- **EarlyStopping:** Interrompe o treinamento se a perda de validação não melhorar após um número definido de épocas, restaurando os melhores pesos.
- **ReduceLROnPlateau:** Reduz a taxa de aprendizado se a perda de validação estagnar, ajudando a afinar o ajuste dos pesos.

O otimizador utilizado foi o *Adam*, e a função de perda foi a *categorical crossentropy*, adequada para problemas de classificação multiclasse.

Visualização dos Resultados. Funções auxiliares, como `plot_losses` e `plot_confusion_matrix`, foram desenvolvidas para monitorar a evolução da perda durante o treinamento e para analisar a matriz de confusão, respectivamente, possibilitando uma análise detalhada dos erros de classificação.

5 Experimentos

5.1 Configuração Experimental

- **Ambiente de Execução:** O sistema foi treinado e avaliado em uma máquina equipada com processador Ryzen 5 3600, placa de vídeo RTX 3050 e 16GB de memória RAM, possibilitando o processamento paralelo necessário para o treinamento dos modelos.
- **Datasets Utilizados:** Foram utilizadas bases de dados previamente validadas, como o LIBRAS-UFOP, que asseguram a diversidade dos sinais e condições de captura. Onde selecionamos apenas os dados das categorias 1 e 2. Sendo primeiramente treinado e validado com a categoria 2 e posteriormente refeito com as categorias 1 e 2 juntas.

5.2 Métricas de Avaliação

A performance do modelo foi avaliada utilizando as seguintes métricas:

- **Acurácia:** Proporção de sinais corretamente classificados.
- **Precisão (Precision) – Macro:** Capacidade do modelo de minimizar falsos positivos.
- **Revocação (Recall) – Macro:** Eficiência na identificação correta dos sinais.
- **F1-Score – Macro:** Média harmônica entre precisão e revocação, refletindo a robustez geral do modelo.

6 Resultados

Após o treinamento, analisamos o desempenho por meio de métricas padrão de classificação. A Tabela 1 sintetiza as métricas obtidas no conjunto de teste para o modelo Transformer (com derivadas temporais):

Tabela 1: Comparação de métricas de reconhecimento para o modelo Transformer.

Modelo (com derivada)	**Acurácia**	**Precisão (macro)**	**Revocação (macro)**	**F1 (macro)**
Transformer (apenas categoria 2)	98.62%	98.48%	98.68%	98.56%
Transformer (categoria 1 e 2)	84.12%	88%	85%	83%

Os experimentos com validação cruzada K-Fold evidenciaram resultados médios superiores a 98% de acurácia com dados apenas da categoria 2. Já com dados das categorias 1 e 2, os resultados médios foram superiores a 84% de acurácia, demonstrando a eficácia da abordagem Transformer na tarefa de reconhecimento de sinais.

Observa-se que a inclusão das derivadas temporais enriquece a representação dos movimentos, possibilitando uma discriminação fina entre gestos com variações sutis. Adicionalmente, a matriz de confusão (Figura 3 e 4) permitiu identificar as classes com maiores desafios de classificação, corroborando achados anteriores de que gestos similares exigem estratégias robustas para capturar diferenças discretas.

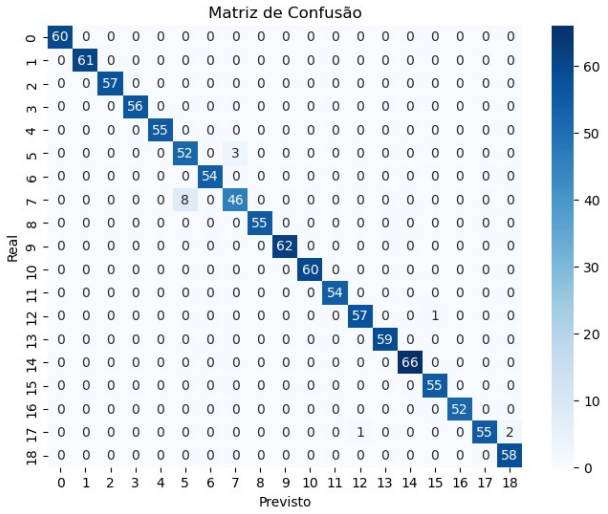

Figura 3: Matriz de confusão para dados de categoria 2.

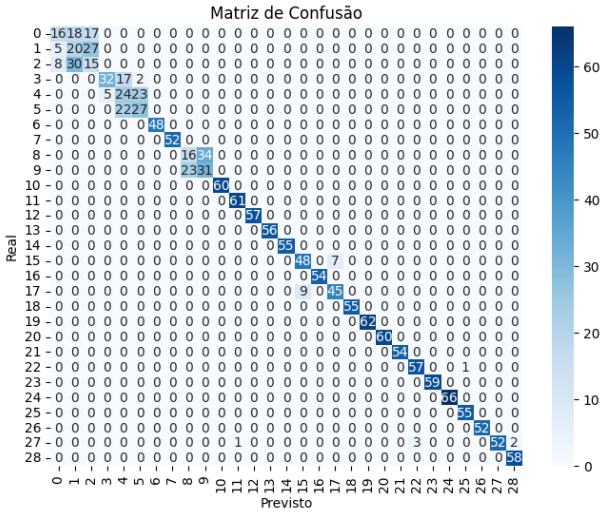

Figura 4: Matriz de confusão para dados de categoria 1 e 2.

7 Comparação com Trabalho Original

Em comparação com abordagens anteriores, especialmente os estudos que utilizaram dados multimodais (RGB, profundidade e esqueleto), o sistema proposto apresenta diferenças significativas tanto em termos de arquitetura quanto de desempenho. Destacam-se:

7.1 Vantagens

- **Simplicidade da Arquitetura:** Ao focar exclusivamente em dados de esqueleto, a complexidade computacional é reduzida, eliminando a necessidade de processar imagens RGB e dados de profundidade.

- **Mecanismo de Atenção Eficiente:** O uso do Transformer com Multi-Head Attention permite capturar relações de longo alcance e identificar os frames mais relevantes, sem a limitação sequencial dos LSTMs.

- **Processamento Paralelo:** A arquitetura Transformer possibilita treinamento mais rápido e eficiente devido à capacidade de processamento paralelo.

- **Representação Dinâmica:** A inclusão de derivadas temporais enriquece a representação, aumentando a sensibilidade do modelo a variações sutis nos gestos.

7.2 Desafios e Limitações

- **Dependência da Qualidade dos Dados:** A eficácia do sistema depende fortemente da precisão na captura das juntas. Ruídos e imprecisões podem afetar o desempenho, exigindo técnicas robustas de pré-processamento.
- **Complexidade do Ajuste dos Hiperparâmetros:** A definição adequada dos parâmetros do Transformer (como número de cabeças e dimensão das chaves) é crucial e demanda ajustes finos para maximizar a performance.

7.3 Pontos de Melhoria Futura

- **Fusão Multimodal:** Investigar a integração de informações de dados RGB e profundidade para complementar a abordagem baseada em esqueleto.
- **Estratégias Semi-Supervisionadas:** Explorar métodos que reduzam a dependência de grandes volumes de dados anotados, ampliando a aplicabilidade do modelo.
- **Aprimoramento do Mecanismo de Atenção:** Refinar os componentes do Transformer para aumentar a sensibilidade às variações temporais e espaciais sutis presentes em gestos semelhantes.

8 Conclusão

Este trabalho reproduz e adapta a proposta de Cerna et al. (2021) para o reconhecimento de sinais da LIBRAS, utilizando exclusivamente dados de esqueleto e implementando uma arquitetura Transformer com Multi-Head Attention. A pipeline, que abrange desde o pré-processamento e extração de características (incluindo derivadas temporais) até a validação cruzada K-Fold, demonstrou alta eficácia, alcançando uma acurácia média de 98.62% e 84.13% com métricas robustas de precisão, revocação e F1-score. A comparação com abordagens anteriores evidencia as vantagens do Transformer, especialmente em termos de processamento paralelo e capacidade de capturar relações temporais de longo alcance, apontando direções promissoras para futuras melhorias e integração multimodal.

Referências

Alves, C. E. G. R., Boldt, F. d. A., & Paixão, T. M. (2024). Enhancing brazilian sign language recognition through skeleton image representation. *arXiv preprint arXiv:2404.19148*. Retrieved from https://arxiv.org/abs/2404.19148

Cerna, L. R., Cardenas, E. E., Miranda, D. G., Menotti, D., & Camara-Chavez, G. (2021). A multimodal libras-ufop brazilian sign language dataset of minimal pairs using a microsoft kinect sensor. *Expert Systems with Applications*, *167*, 114179.

Chen, H., Wang, J., Guo, Z., Li, J., Zhou, D., Wu, B., ... Heng, P.-A. (2024). Signvtcl: multi-modal continuous sign language recognition enhanced by visual-textual contrastive learning. *arXiv preprint arXiv:2401.11847*.

De Souza, F., Camara-Chavez, G., do Valle, E., & De A Araujo, A. (2010). Violence detection in video using spatio-temporal features. In *Graphics, patterns and images (sibgrapi), 2010 23rd sibgrapi conference on* (p. 224-230).

Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2969–2978).

Hu, L., Gao, L., Liu, Z., & Feng, W. (2024). Dynamic spatial-temporal aggregation for skeleton-aware sign language recognition. *ArXiv*, *abs/2403.12519*. Retrieved from https://api.semanticscholar.org/CorpusID:268531538

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton aware multi-modal sign language recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3413–3423).

OPAS/OMS. (2021). *Oms publica primeiro relatório global sobre inteligência artificial na saúde e seis princípios orientadores para sua concepção e uso.* Retrieved from https://www.paho.org/pt/noticias/28-6-2021-oms-publica-primeiro-relatorio-global-sobre-inteligencia-artificial-na-saude-e (Acessado em: 22 jan. 2025)

Özdemir, O., Baytaş, İ. M., & Akarun, L. (2023). Multi-cue temporal modeling for skeleton-based sign language recognition. *Frontiers in Neuroscience*, *17*, 1148191.

Plaza, J. V., de la Escalera Hueso, A., & Moreno, J. M. A. (2024). Transformer autorregresivo de grafos esqueléticos. *Jornadas de Automática*(45).

Souza, M. F. N. S. d., Araújo, A. M. B., Sandes, L. F. F., Freitas, D. A., Soares, W. D., Vianna, R. S. d. M., & Sousa, Á. A. D. d. (2017). Principais dificuldades e obstáculos enfrentados pela comunidade surda no acesso à saúde: uma revisão integrativa de literatura. *Revista Cefac*, *19*(3), 395–405.

Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., & Tang, H. (2020). Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 1203–1211.

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE Multimedia*, *19*(2), 4–10. doi: 10.1109/MMUL.2012.24