Reconhecimento da Língua Brasileiras de Sinais (LIBRAS) em tempo real

Guilherme Soares , Wesna Simone , Lucas Yagui

Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
CEP 13083-852 – Campinas, SP, Brasil

{g217241,w225843,l240211}@dac.unicamp.br

Resumo – O relacionamento social ocorre através de uma importante ferramenta: a comunicação. Por meio desta, é possível desenvolver o aprendizado e entender as necessidades que envolvem a sociedade/individuo. No entanto, a comunicação pode ser falha quando um dos seus principais canais (fala e audição) é comprometido. Pensando nisso, o presente trabalho busca desenvolver um protótipo que identifica a língua brasileira de sinais (LIBRAS) para apenas nove letras através do uso de metodologias de aprendizado de máquina. Assim, a primeira parte do projeto tratará da preparação do banco de dados bem como seu treinamento e por fim a predição do modelo feita em tempo real com a utilização de uma *webcam*.

Palavras-chave – LIBRAS, Linguagem de sinais, Aprendizado de máquina, Comunicação, Gestos, Reconhecimento.

1. Introdução

Libras é a sigla para língua brasileira de sinais, uma língua de modalidade gestual-visual em que é possível se comunicar por meio de gestos, expressões faciais e corporais. Desde 24 de abril de 2002, Libras se tornou uma língua oficial brasileira.

Essa linguagem é um meio de comunicação muito importante para a comunidade não só de surdos e mudos, mas também para qualquer pessoa que irá interagir com esses indivíduos. Exemplo disso é o aumento no aparecimento de intérpretes em vídeos e transmissões ao vivo. No entanto, essa interpretação nem sempre é de fácil acesso para a população.

Dessa forma, visando promover a acessibilidade pensando no uso constante de meios de comunicação durante o período de pandemia foi desenvolvido um protótipo que reconhece nove letras do alfabeto da língua brasileira de sinais (C, E, F, I, M, N, O, R, S) utilizando aplicações de aprendizado de maquina e redes convolucionais. Além disso, para aproximar o projeto de um possível uso cotidiano a predição do modelo foi desenvolvida através da coleta de imagens retiradas da *webcam* em tempo real.

É importante destacar que o trabalho realizado foi baseado em outros dois projetos de linguagem de sinais [1, 2] com alterações principalmente no treinamento da rede, manipulação do banco de dados e predição das imagens.

2. Proposta

A proposta do trabalho foi dividida em duas etapas: A primeira consistiu em desenvolver uma rede neural convolucional que fosse capaz de classificar corretamente nove letras do alfabeto de Libras. Os gestos são mostrados na Figura 1. Para isso foi utilizado um banco de treino com 18953 imagens 64x64 e um conjunto de teste de 6257 com as mesmas dimensões. Além disso, o número de letras foi reduzido, pois a proposta inicial tinha por objetivo construir o próprio banco de dados. As letras foram escolhidas de modo a formar o nome/apelido dos nossos professores: "CEFINO E ROMIS". Lembrando, que não há a presença da letra H no banco de dados, pois essa exige movimentação que inicialmente não haverá suporte no projeto.

Já a segunda etapa tinha por objetivo captar imagens em tempo real da *webcam* (frames) e com isso predizer o modelo.

2.1. Dataset

Inicialmente foi projetado a utilização de 450 imagens de treino e 150 de teste por classe em fundo branco. Para tanto, foi adotado um modelo conforme aplicativo de captura já disponibilizado na plataforma github [2], consistindo de imagens de tamanho 64x64 pixels com três canais (RGB). Para efeitos de treinamento, o conjunto de imagens foi convertido em arquivo CSV, contendo rotulação e vetores correspondentes.

Entretanto, ao tentar realizar o treinamento da rede utilizando essa configuração, não foram obtidos resultados satisfatórios. Em outras palavras, a rede não atingiu acurácia suficiente para que fosse possível reconhecer propriamente os gestos - cerca de 50%.

Com o objetivo de obter parâmetros de comparação em relação à acurácia do modelo, foi utilizado um dataset referente a Linguagem de Sinais Americana [3] (imagens em 28x28x3). Embora o resultado tenha atingido 90% de acurácia, não apresentou reconhecimento preciso nos testes realizados, visto que as imagens capturadas se mostraram distantes das utilizadas pelo banco de dados.

Por fim, baseando-se na diferença de acurácia apresentada, percebeu-se que seria necessário uma quantidade maior de dados para treinamento. Com isso, para ampliação do dataset feito pelo grupo, foi anexado ao

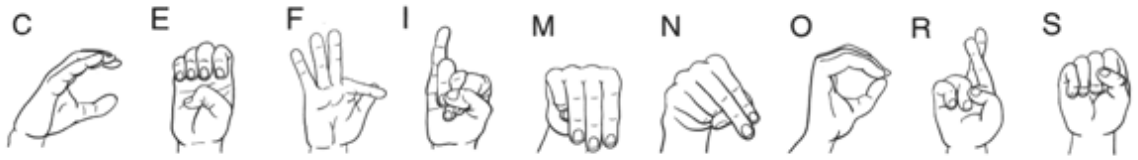

Figura 1. Modelo de gestos das letras utilizadas no dataset conforme alfabeto de Libras.

projeto o banco de dados disponibilizado em [2], que seguia o padrão de imagens adotado desde o início, apenas sendo necessário o processamento para convertê-lo ao modelo CSV. A distribuição por classe resultante é demonstrada nas Figuras 2 e 3.

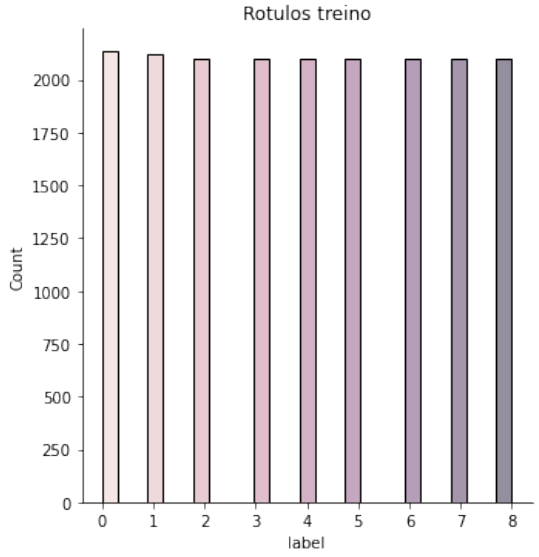

Figura 2. Distribuição por classe do dataset de treino final. Os rótulos se referem a: 0 - C, 1 - E, 2 - F, 3 - I, 4 - M, 5 - N, 6 - O, 7 - R, 8 - S.

Nota-se através dos histogramas que o banco de dados está bem distribuído, apresentando algumas variações no conjunto de dados de testes, fato que não demonstrou prejuízo no desempenho do modelo durante verificação.

2.2. Rede Convolucional

Após planejamento a cerca do banco de dados, foi utilizada a arquitetura apresentada na Figura 4. Optou-se por uma rede densa, com o objetivo de extrair mais características únicas de cada classe, dado que o conteúdo das imagens foram julgados parecidos.

A partir disso, foi realizado o treinamento, cujos resultados são ilustrados nas Figuras 5 e 6. É possível perceber que a acurácia do modelo esteve próxima aos 99%, enquanto o grupo de teste se manteve próximo aos 95%. Além disso, é evidente que a acurácia de treino

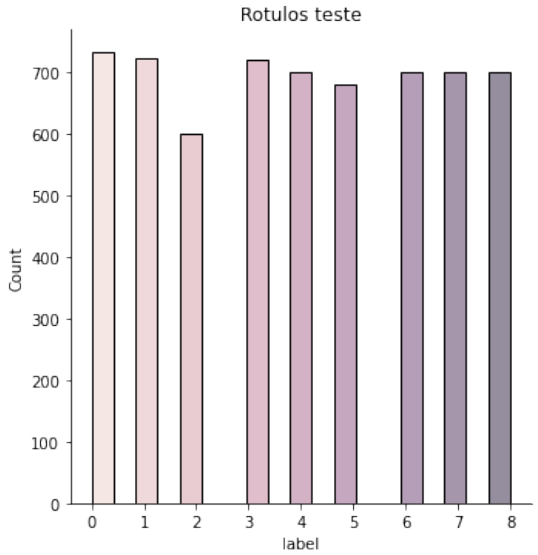

Figura 3. Distribuição por classe do dataset de teste final. Os rótulos se referem a: 0 - C, 1 - E, 2 - F, 3 - I, 4 - M, 5 - N, 6 - O, 7 - R, 8 - S.

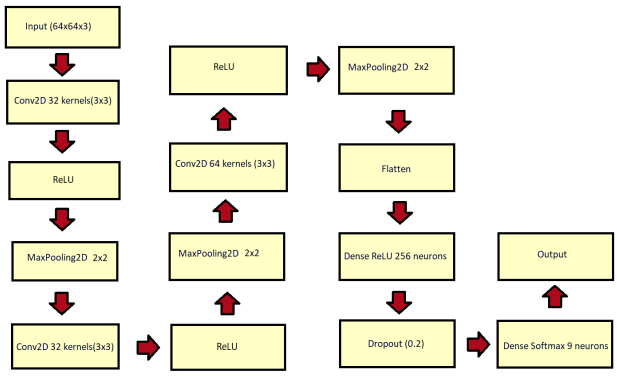

Figura 4. Arquitetura da rede convolucional do modelo para reconhecimento de Libras em tempo real.

apresentou baixa variação, enquanto que a de teste oscilou ao longo de todas as épocas.

Em relação ao gráfico de perdas, nota-se um grande aumento do custo por época no grupo de teste, sinalizando possibilidade de *overfitting*. Isso demonstra que uma menor quantidade de épocas seria suficiente.

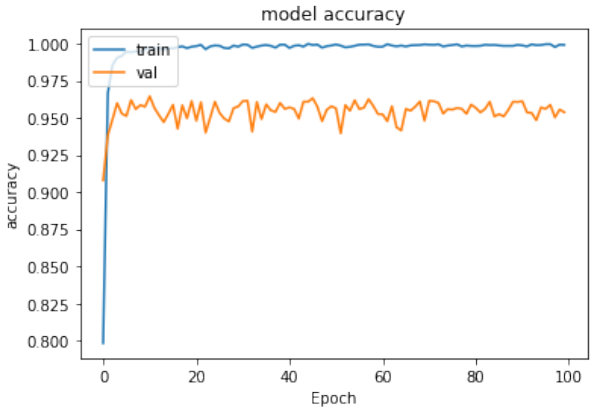

Figura 5. Gráfico de acurácia por época do modelo proposto. Em azul, o grupo de treino, e em laranja, o de teste.

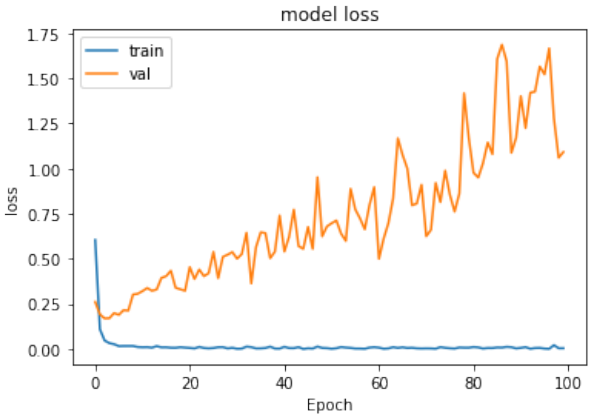

Figura 6. Gráfico de perdas por época do modelo proposto. Em azul, o grupo de treino, e em laranja, o de teste.

Em seguida, para análise da classificação da rede treinada, foi gerada a matriz de confusão apresentada na Figura 7. É perceptível que as classes 'M' e 'N' apresentaram grandes equívocos entre elas para a rede, visto a maior quantidade de falsos positivos em comparação com outras classes. Isso possivelmente se deve à semelhança dos gestos, como mostrado na Figura 1.

Nesse contexto,realizou-se tentativas de predição com o grupo de testes para verificação de desempenho. Com isso, gerou-se a Figura 8 com exemplos de sucesso na predição.

2.3. Predição em tempo real

Após treinamento, foi necessário desenvolver formas de obtenção de imagens provenientes da *webcam* do usuário. É importante destacar que o recolhimento de dados nesse estágio é feito através da captura de frames da câmera, que normalmente trabalha em 30 frames por segundo (FPS). Assim, foi delimitada uma região da câmera para auxiliar o posicionamento correto das mãos, onde será recolhida a foto para processamento.

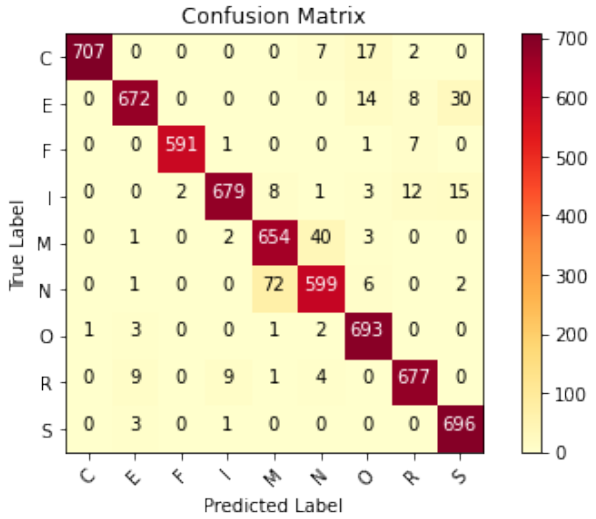

Figura 7. Matriz de confusão do modelo proposto. Nota-se um maior equívoco entre as classes 'M' e 'N'.

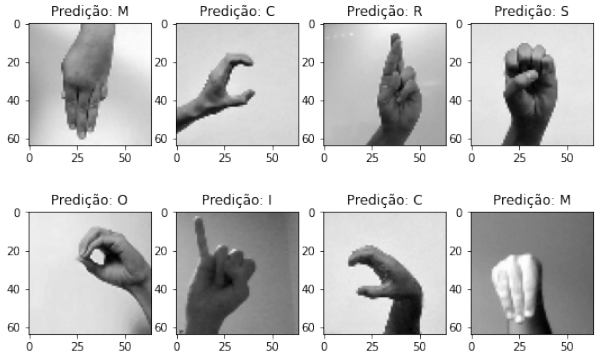

Figura 8. Conjunto de imagens em que o modelo realizou com sucesso a predição.

Como a resolução de captura é relativamente maior que o tamanho das imagens utilizadas no projeto, mostrou-se necessário redimensionamento da imagem, além da mudança de paleta de cores para a escala de cinza, facilitando a padronização do reconhecimento. Dessa forma, ocorreu a transformação do frame capturado, ao apertar a barra de espaço, em vetor normalizado, enviando-o para classificação do modelo.

3. Resultados

As Figuras 9, 10 e 11 foram obtidas utilizando o protótipo desenvolvido. Nota-se que as predições foram realizadas pelos autores do projeto em fundo branco. As janelas dispostas envolvem a leitura da câmera com o retângulo de interesse em destaque, a imagem cortada na dimensão do retângulo, sua réplica em escala de cinza e a janela de predição.

Um vídeo demonstrando os resultados obtidos pode ser visualizado clicando **aqui**.

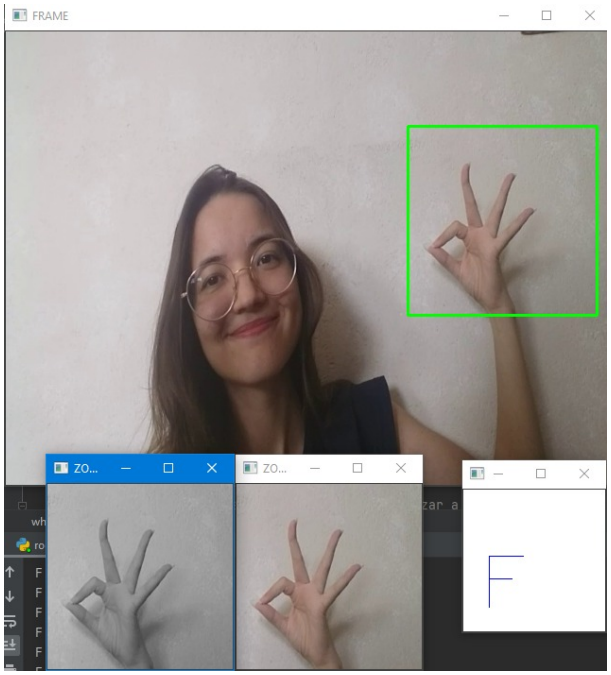

Figura 9. Resultado de predição da letra F. Imagem: autor Wesna Simone.

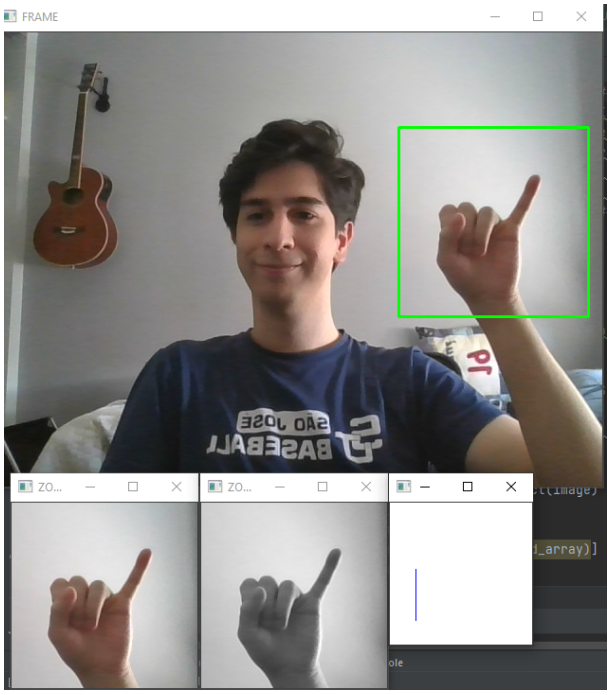

Figura 10. Resultado de predição da letra I. Imagem: autor Lucas Yagui.

4. Conclusões

Neste trabalho foi investigado a funcionalidade de um sistema de reconhecimento de Libras em imagens provenientes da *webcam* de usuários. Foram abordados formas de geração de banco de dados em CSV, manipulação de dados em tempo real, sistemas introdutórios de processamento de imagens e aprendizado de máquina utilizando redes neurais convolucionais.

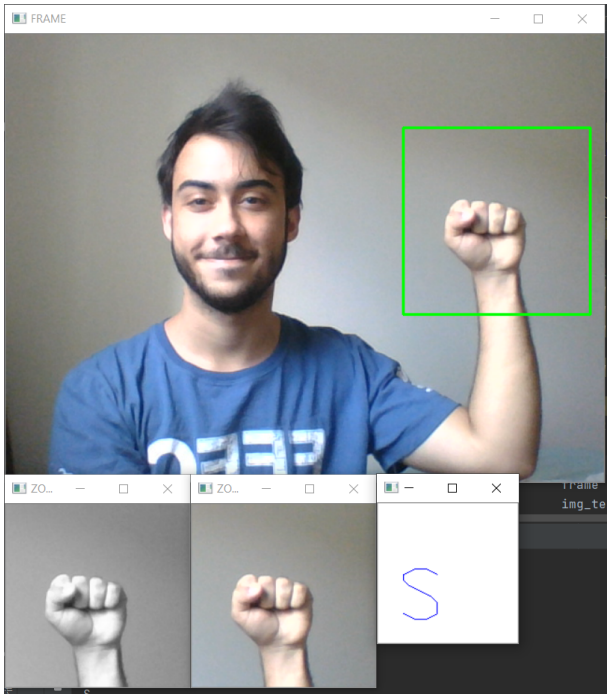

Figura 11. Resultado de predição da letra S. Imagem: autor Guilherme Soares.

Os resultados do modelo criado se mostraram bastante satisfatórios, visto aproveitamento de 95% para acurácia de validação. Além disso, nos testes em tempo real, os gestos foram captados de forma rápida e precisa, ocorrendo poucos equívocos, que se concentraram principalmente nas letras 'M' e 'N' devido às suas semelhanças.

Visando aperfeiçoamento do protótipo desenvolvido, possíveis melhorias para o futuro envolveriam a expansão do banco de dados para outras letras, bem como inserção de expressões características da língua - o que levaria a cada vez maior proximidade com o uso cotidiano, além de melhorias nas técnicas de processamento de imagem, com objetivo de permitir a utilização do modelo em planos de fundo diferentes, não restrito ao fundo branco. Ademais, a inclusão de algoritmos de detecção de mãos seriam benéficos, de modo a permitir o posicionamento em qualquer região da câmera.

De modo geral, o modelo mostrou-se surpreendente para os autores do trabalho, instigando aprofundamento do conhecimento nas áreas abordadas.

Referências

[1] Brenner Heintz. Training a Neural Network to Detect Gestures with OpenCV in Python. (acessado em 24/12/2020).

[2] Lucas Lacerda. Redes Neurais Convolucionais - LIBRAS. *Github*, Janeiro 2019.

[3] Usuário tecperson. Sign Language MNIST. (acessado em 25/12/2020).