Trabalho Final

Integrantes do grupo:
Diego Roberto Das Chagas - 2021019718

Felippe Veloso Marinho - 2021072260

1. Descrição do problema a ser resolvido

Nesse trabalho, será abordada a aplicação de redes neurais no contexto de Reconhecimento Contínuo de Linguagem de Sinais Baseado na Visão (CSRL).

O Reconhecimento Contínuo de Língua de Sinais baseado na Visão (CSLR) é uma tecnologia que utiliza dados visuais, como vídeos, imagens, para identificar e interpretar automaticamente sinais de língua de sinais em tempo real. Ao contrário do reconhecimento isolado de sinais, onde cada sinal é analisado de forma independente, o CSLR lida com fluxos contínuos de sinais, permitindo a tradução de frases inteiras e conversas em linguagem de sinais.

O principal objetivo do CSLR é traduzir automaticamente a língua de sinais para texto ou fala, facilitando a comunicação entre pessoas surdas e ouvintes. Para implementar o CSLR, são usadas redes neurais avançadas, o que inclui, por exemplo, as Redes Neurais Convolucionais (CNNs), que serão utilizadas nesse trabalho.

Um dos principais problemas enfrentados no uso de redes neurais para interpretação de línguas de sinais é o overfitting. A ocorrência de overfitting nesse contexto é comum, devido à alguns fatores como diferenças individuais na reprodução dos sinais, complexidades dos modelos, dados limitados, etc. Existe, por exemplo, uma grande variabilidade na forma como diferentes pessoas executam os mesmos sinais. Essa variabilidade pode ser difícil de capturar com precisão, levando o modelo a ajustar-se demais aos sinais específicos de indivíduos presentes no conjunto de treinamento. Também há uma quantidade limitada de dados disponíveis para treinar os modelos. Com menos dados, as redes neurais são mais propensas a memorizar exemplos específicos em vez de aprender padrões generalizáveis.

Serão propostas, nesse trabalho, algumas técnicas e formas de tentar amenizar o problema do overfitting no contexto apresentado.

2. Revisão Bibliográfica

Atualmente, diferentes métodos são aplicados na tentativa de mitigar o overfitting no contexto supracitado. Na tentativa de amenizar a baixa disponibilidade de dados para treinamento temos, por exemplo, o “Data Augmentation”. Esse método envolve a aplicação de transformações aleatórias aos dados de entrada, como rotação, zoom, e mudanças de iluminação. Isso cria variações nos dados de treinamento, aumentando a robustez do modelo e ajudando-o a generalizar melhor.

Outra técnica para atuar contra o overfitting é Regularização. Essa técnica tem o intuito de prevenir que o modelo se ajuste excessivamente aos dados de treinamento. Uma das formas de aplicar a regularização é através do método

"Dropout", que ajuda a prevenir o overfitting ao desligar aleatoriamente unidades (neurônios) na rede durante o treinamento. Esse processo impede que a rede se torne excessivamente dependente de neurônios específicos, promovendo a independência e robustez dos neurônios.

3. Metodologia utilizada

A metodologia selecionada para solução do problema proposto foi o uso de Redes Neurais Convolucionais no contexto de CSLR, juntamente com a aplicação do "Visual Alignment Constraint (VAC)".

3.1 Redes Neurais Convolucionais

As redes neurais convolucionais (Convolutional neural network ou CNNs) são um subconjunto do aprendizado de máquina utilizadas com mais frequência para tarefas de classificação e visão computacional. Essas redes oferecem uma abordagem mais dimensionável para tarefas de classificação de imagens e reconhecimento de objetos. Sendo assim, as CNNs se diferenciam das demais redes neurais por seu desempenho superior com entradas de imagens, fala ou sinais de aúdio.

Essas redes possuem três principais tipos de camadas: camada convolucional, camada de agrupamento e camada totalmente conectada.
Basicamente, a camada convolucional é a principal responsável pelos cálculos, utilizando filtros que passam pela imagem para identificar características específicas, como bordas ou texturas. Esses filtros geram mapas de feições que ajudam a rede a reconhecer padrões. As camadas de agrupamento reduzem a dimensionalidade dos dados, reduzindo o número de parâmetros nas entradas. Semelhante à camada convolucional, a operação de agrupamento varre um filtro por toda a entrada, a diferença é que este filtro não tem nenhum peso. Em vez disso, o kernel aplica uma função de agregação aos valores dentro do campo receptivo, preenchendo a matriz de saída. Esse processo, por consequência, ajuda a diminuir a complexidade e evitar o superajuste (overfitting). A camada totalmente conectada executa a tarefa de classificação baseada nas feições extraídas através das camadas anteriores e seus diferentes filtros. Ela conecta cada nó a todos os nós da camada anterior, realizando a classificação final das características extraídas.

Redes neurais convolucionais são uma boa escolha de modelo a utilizar em reconhecimento contínuo de língua de sinais devido à sua capacidade de extrair automaticamente características visuais relevantes das sequências de vídeo. As CNNs capturam padrões espaciais e temporais nas imagens, permitindo a identificação de sinais com alta precisão. Elas são eficazes na detecção de variações nas expressões faciais e nos movimentos das mãos, elementos essenciais para interpretar a língua de sinais.

3.2 Visual Alignment Constraint (VAC)

O Visual Alignment Constraint (VAC) é um método que visa reduz o overfitting

em CSRLs, melhorando a robustez e a generalização das redes neurais convolucionais (CNNs). Esse método introduz dois tipos de perdas nas redes neurais convolucionais: a perda de características visuais e a perda de alinhamento temporal.

A perda de características visuais incentiva a rede neural a aprender representações visuais mais robustas e consistentes e não apenas memorizações específicas, forçando a rede a focar em características que sejam mais relevantes e generalizáveis. Já a perda de alinhamento temporal tem o intuito de assegurar que as características aprendidas pela rede estejam alinhadas corretamente ao decorrer das diferentes etapas temporais. Isso resulta em uma interpretação mais coerente de uma sequência de sinais, o que é essencial para lidar com a evidente variabilidade das línguas de sinais.

Essas medidas adotadas pelo método VAC resultam em na regularização do treinamento e em uma melhor consistência temporal, melhorando a generalização dos dados e evitando que o modelo memorize dados de treinamento e, portanto, amenizando o recorrente problema de overfitting.

4. Descrição dos dados

Os datasets utilizados contêm imagens do alfabeto em ASL (Linguagem de Sinais Americana), onde cada imagem está associada a um label que varia de 1 a 29 para representar as letras do alfabeto ASL. Ambos os datasets podem ser encontrados na plataforma Hugging Face, sendo eles:

- ASL Sign Languages Alphabets v02
- ASL Sign Languages Alphabets v03

A estrutura de cada dataset é a seguinte:

- **Imagens:** Contêm fotos de sinais de mão representando cada letra do alfabeto ASL.
- **Labels:** Um número entre 1 e 29 que corresponde à letra representada pela imagem.

4.1 Motivação para a Junção dos Datasets

A junção de ambos os datasets foi motivada pela necessidade de aumentar a quantidade de dados de treinamento. Em problemas de aprendizado de máquina, especialmente com redes neurais profundas, a quantidade de dados disponível pode impactar significativamente o desempenho do modelo. Mais dados geralmente resultam em um modelo mais robusto e menos suscetível ao overfitting. Ao combinar os dois datasets, conseguimos aumentar a diversidade das imagens e garantir uma melhor cobertura das possíveis variações nos sinais de mão, melhorando assim a capacidade do modelo de generalizar para novos dados.

5. Experimentos e Resultados

Inicialmente, os dados foram divididos em conjuntos de treinamento e teste, utilizando uma proporção de 70% para treinamento e 30% para testes.

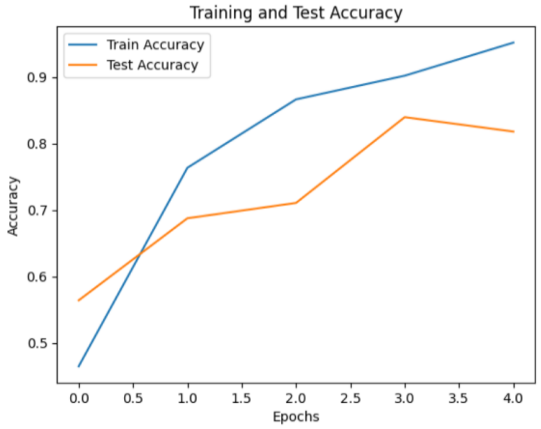
Os experimentos foram conduzidos com o objetivo de treinar o modelo com e sem a restrição de alinhamento visual para mitigar o overfitting.

As métricas utilizadas incluíram a verificação da acurácia nos conjuntos de treinamento e teste, além da análise das curvas de aprendizado com validação ao longo das épocas.

A análise foi inspirada nas sugestões presentes no paper e focou na redução da diferença entre as perdas de treinamento e validação, indicativa de menor overfitting. A precisão no conjunto de treinamento também foi comparada com a precisão no conjunto de teste: uma grande disparidade entre essas precisões sugere overfitting.

Primeiro foram gerados os gráficos para o modelo sem a utilização das técnicas de Restrição de Alinhamento Visual. Os resultados para os gráficos de Training and Test Loss e Training and Test Accuracy foram os seguintes:

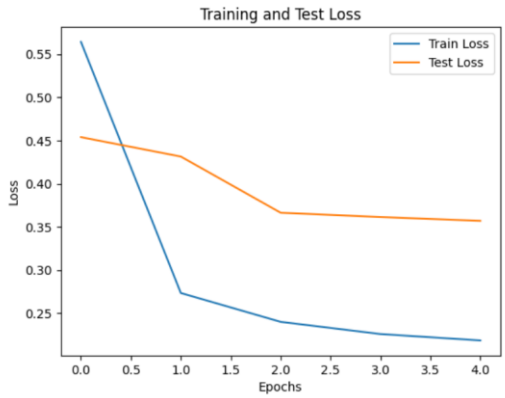
Durante as épocas, a perda nos dados de teste (Test Loss) reduziu-se mais rapidamente do que nos dados de treinamento (Train Loss). Isso sugere que o modelo está generalizando melhor para dados novos, indicando uma possível redução no overfitting mesmo sem a utilização de VACs. Apesar disso, em determinado número de épocas a curva com os dados de teste apresentou um maior número de perdas. O que pode representar um indício de overfitting nessa quantidade de épocas.



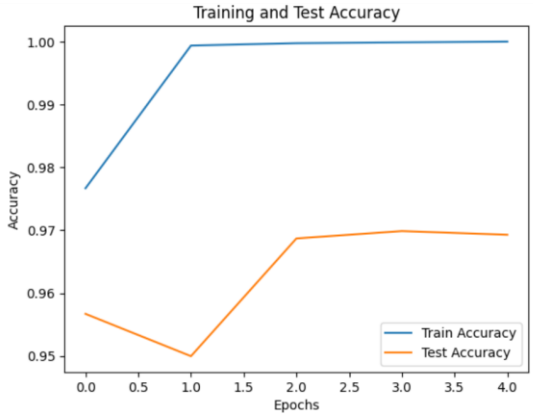
A acurácia tanto no treinamento quanto no teste aumentou ao longo do tempo. Esse aumento indica que o modelo está melhorando sua capacidade de fazer previsões corretas, tanto nos dados usados para treinamento quanto nos dados novos (teste). Novamente, temos um resultado que reflete uma piora de precisão, o que alerta novamente para a possibilidade de overfitting ao aumentar o número de épocas.

No segundo teste utilizamos o modelo com Restrição de Alinhamento Visual com a intenção de aprimorar a capacidade de generalização do extrator visual restringindo o espaço de

recursos com a supervisão do alinhamento.



A utilização das técnicas de VAC parece ter contribuído para uma redução adicional no overfitting, como indicado pela menor diferença entre as perdas de treinamento e teste.



A proximidade entre os dados de acurácia de treinamento e teste indica que o modelo está generalizando melhor. Quando as curvas de acurácia de treinamento e teste estão mais próximas, isso é um indicativo positivo de que o modelo está se comportando bem em dados não vistos durante o treinamento.

Após a aplicação de técnicas de VAC para mitigação do overfitting, foi observado que as perdas tanto no treinamento quanto no teste são maiores, o que pode indicar uma redução adicional na disparidade entre os conjuntos. Isso sugere que o modelo está generalizando melhor para novos dados. Além disso, a proximidade entre as curvas de acurácia de treinamento e teste também é evidente, o que reforça a melhoria na capacidade do modelo em lidar com dados não vistos durante o treinamento.

6. Conclusão

Este trabalho explorou o Reconhecimento Contínuo de Linguagem de Sinais Baseado na Visão (CSLR) usando Redes Neurais Convolucionais (CNNs), com foco na mitigação do overfitting. Estratégias como Data Augmentation, Regularização (especialmente Dropout) e o método Visual Alignment Constraint (VAC) foram empregadas para melhorar a generalização do modelo. Os resultados indicaram que o VAC reduziu significativamente o overfitting, conforme evidenciado pela menor disparidade entre as perdas de treinamento e teste, além da proximidade das curvas de acurácia de treinamento e teste.

7. Link da Apresentação do Trabalho

Link para apresentação: https://youtu.be/Obznll5bBJY

8. Referências

[1] Yuecong Min1,2 , Aiming Hao1,2 , Xiujuan Chai3 , Xilin Chen1,2 Visual Alignment Constraint for Continuous Sign Language Recognition

[2] Introdução ao PyTorch | Redes Neurais | Primeiros passos com Pytorch | Deep Learning #1| Programação Dinâmica

[3] Redes Neurais Convolucionais com PyTorch | Visão Computacional | Deep Learning #2 | Programação Dinâmica

[4] resnet18 | Documentação | PyTorch.org

[5] Projeto Captar-Libras | VerLab

[6] Asl_sign_languages_alphabets_v03

[7] Asl_sign_languages_alphabets_v02