1. Descrição do Problema

Treinar um modelo capaz de identificar as 5 classes especificadas em um vídeo fornecido:

- Aplicar Batom
- Cortar Cabelo
- Secar Cabelo com Secador de Mão
- Aplicar Maquiagem nos Olhos
- Escovar os Dentes

Validar o modelo e aplicar a inferência no vídeo de exemplo fornecido
(`data/Raw/test-vid/test-actions-vid.mp4`).

A saída deve conter:

- Um arquivo de vídeo com informações sobre a ação atual sendo realizada, sobrepostas no vídeo.
- Um documento que descreve as ações na ordem, com tempo de início, tempo de término e duração.

2. Conjuntos de Dados

2.1 UCF101

Visão Geral

O UCF101 é um conjunto de dados de reconhecimento de ações realistas, coletado do YouTube, contendo 101 categorias de ações. Este conjunto é uma extensão do UCF50 e consiste em 13.320 clipes de vídeo, classificados em 101 categorias. Essas categorias podem ser agrupadas em cinco tipos:

1. Interação Humano-Objeto
2. Apenas Movimento Corporal
3. Interação Humano-Humano
4. Tocando Instrumentos Musicais
5. Esportes

Todos os vídeos foram coletados do YouTube e possuem uma taxa de quadros fixa de 25 FPS, com resolução de 320 × 240.

Detalhes do Conjunto de Dados

Os vídeos em 101 categorias de ações são agrupados em 25 grupos, onde cada grupo pode consistir de 4 a 7 vídeos de uma ação. Os vídeos do mesmo grupo podem compartilhar algumas características comuns, como fundo semelhante, ponto de vista semelhante, etc. (IMPORTANTE: considerar isso ao separar treino e teste)

Note: http://www.thumos.info/ possui um dataset de "background", se trata de videos que necessariamente não contém nenhuma das 101 atividades indicadas pelas 101 classes do dataset.

Classes Disponíveis

O conjunto de dados UCF101 inclui as seguintes classes que correspondem às ações de interesse:

- **Aplicar Batom**
- **Cortar Cabelo**
- **Secar Cabelo com Secador de Mão**
- **Aplicar Maquiagem nos Olhos**
- **Escovar os Dentes**

Este conjunto de dados também possui classes parecidas com as 5 de interesses como "Head Massage". Podem ser usadas como background, para tentar ajudar o modelo a aprender melhor as classes de interesse.

2.2 Kinetics

Visão Geral

O Kinetics é uma coleção de conjuntos de dados de grande escala para reconhecimento de ações humanas em vídeos. Dependendo da versão, o Kinetics abrange entre 400 e 700 classes de ações, com pelo menos 400 vídeos por classe. Cada vídeo tem aproximadamente 10 segundos e é extraído do YouTube. As ações incluem interações humano-objeto, como tocar instrumentos musicais, e interações humano-humano, como apertar as mãos.

Detalhes do Conjunto de Dados

- **Número de Classes**: 400 (Kinetics-400), 600 (Kinetics-600) ou 700 (Kinetics-700), dependendo da versão.
- **Número de Vídeos**: Aproximadamente 500.000 clipes cobrindo 600 classes de ações humanas, com pelo menos 600 clipes para cada classe de ação.
- **Duração dos Clipes**: Cada clipe dura cerca de 10 segundos.
- **Fonte**: Vídeos coletados do YouTube.

Classes Disponíveis

O Kinetics inclui as seguintes classes que correspondem às ações de interesse, quando considerada a versão mais recente de 700 classes:

- **Aplicar Batom**: "468 putting on lipstick".
- **Cortar Cabelo**: "220 getting a haircut".
- **Secar Cabelo com Secador de Mão**: "38 blowdrying hair".
- **Escovar os Dentes**: "61 brushing teeth".
- **Maquiagem de olho**: "466 putting on eyeliner" (não inclui sombra)

OBS: Dataset comum de se encontrar modelos prétreinados, todavia, grande demais para considerar em um projeto de tão pouco tempo de desenvolvimento. Usarei este para carregar o modelo prétreinado e fazer fine tunning para o UCF101

3. Modelos

3.1 CNN + LSTM

Funcionamento: Combina uma CNN pré-treinada (ex: ResNet) para extrair características de quadros individuais com uma LSTM/GRU para modelar dependências temporais entre os quadros.
Vantagens:

- Simplicidade de implementação.
- Eficiente para vídeos de duração variável.
  Desvantagens:
- Dificuldade em capturar dependências de longo prazo.
- Pode perder nuances temporais complexas.

3.2 3D Convolutional Networks (C3D, I3D, 3D ResNet)

Funcionamento: Aplica convoluções 3D para extrair padrões espaço-temporais diretamente de clipes de vídeo.
Vantagens:

- Captura movimentos de curto prazo de forma nativa.
- Modelos pré-treinados disponíveis (ex: I3D).
  Desvantagens:
- Alto custo computacional devido às operações 3D.
- Menos eficiente para vídeos longos.

3.3 Two-Stream Networks

Funcionamento:

- **Stream espacial**: Processa quadros RGB usando uma CNN tradicional.
- **Stream temporal**: Analisa fluxo óptico para capturar movimento.
- Combina as predições de ambas as streams.
  Vantagens:
- Alta precisão ao integrar aparência e movimento.
- Ideal para ações com movimentos distintos (ex: secar cabelo).
  Desvantagens:
- Requer cálculo prévio de fluxo óptico (ex: RAFT), aumentando o pré-processamento.

3.4 Temporal Segment Networks (TSN)

Funcionamento: Divide o vídeo em segmentos, extrai características de cada um e agrega os resultados (ex: média).
Vantagens:

- Eficiente para vídeos longos.
- Reduz redundância ao amostrar quadros estratégicos.
  Desvantagens:
- Pode subperformar em ações rápidas ou de curta duração.

3.5 SlowFast Networks

Funcionamento:

- **Slow Pathway**: Baixa taxa de quadros para análise espacial detalhada.
- **Fast Pathway**: Alta taxa de quadros para capturar movimento rápido.
- Combina ambas via conexões laterais.
  Vantagens:
- Estado da arte em datasets como Kinetics.
- Balanceia eficiência e precisão.
  Desvantagens:
- Complexidade arquitetural.

3.6 Video Transformers (TimeSformer, ViViT)

Funcionamento: Utiliza mecanismos de atenção para modelar relações espaço-temporais em tokens de vídeo.
Vantagens:

- Excelente em capturar contextos globais.
- Robustez em datasets grandes (ex: Kinetics-700).
  Desvantagens:
- Alto consumo de memória.
- Requer ajustes finos cuidadosos.

4. Planejamento de Implementação

Escolha do Dataset Inicial: UCF101

- **Motivo**:
  - Menor complexidade (13k vídeos vs. 500k no Kinetics).
  - Disponível diretamente via `torchvision.datasets.UCF101` (integração simplificada).
  - Classes específicas já mapeadas (ex: "Applying Lipstick").
- **Estratégia de Fallback**:
  - Se o desempenho for insuficiente, migrar para o **Kinetics-700** (requer download manual ou API customizada).

Ordem de Complexidade dos Modelos

1. **CNN + LSTM**
   - **Custo**: Baixo (CNN 2D + LSTM leve).
   - **Vantagem**: Ideal para validar rapidamente o pipeline (pré-processamento → treino → inferência).
2. **3D CNN (ex: R(2+1)D)**
   - **Custo**: Moderado (operações 3D, mas pode ser acessada pré-treinada no Kinetics).
   - **Comparação com CNN+LSTM**:
     - **Treino**: Mais lento (clipes de 8-16 quadros exigem mais memória).
     - **Precisão**: Geralmente superior em ações com movimento contínuo.
3. **Two-Stream Network (Stream Espacial + Temporal)**
   - **Custo**: Alto (exige cálculo de fluxo óptico).
4. **SlowFast ou Video Transformers**
   - **Custo**: Muito alto.
   - **Vantagem**: Arquiteturas atualmente Estado da arte em problemas do tipo.

Nota: A prioridade é entregar um MVP funcional, mesmo que com limitações. Modelos complexos podem ser explorados em iterações futuras. Devido a isso, no momento será considerado o modelo 3D CNN, devido a facilidade de acessar uma versão pré-treinada no Kinetics, que possa ser refinada com FineTunning dentro do curto período de implementação do projeto.

4. Próximos passos

Listarei abaixo próximos passos para melhorar o projeto que podem ser feitos em iterações futuras.

4.1 Dataset e preprocessing próprios

- Subsituir a utilização das classes de carregamento e preprocessamento da base por soluções próprias customizáveis, melhorando a customização e organização dos dados e garantindo maior liberdade na preparação de bases customizadas, além de possibilidar preprocessamentos próprios. (OBS: vale lembrar que, como utilizei um modelo pré treinado, faz-se necessário utilizar o preprocessamento igual ao que foi utilizado no processo de treinamento que gerou os pesos em questão)

4.2 Augmentation/Balanceamento da base

- Testar técnicas de augmentação e balanceamento para base para garantir um treinamento mais eficiente.
- Há também a possibilidade de definir pesos maiores para classes mais difíceis com o objetivo de direcionar melhor o treinamento do modelo.
- Além disso, é possível criar uma nova base customizada a aprtir de combinações de classes de diferentes bases (ex: juntar kinetics com UCF101) para garantir uma maior diversidade de bases e vídeos.

4.3 Validar demais arquiteturas/estratégias

Devido ao tempo do projeto não foi possível validar as demais estratégias que são estado da arte em problemas de detecção de ativade humana como **SlowFast** ou **Video Transformers**. Espera-se que modelos estado da arte consigam resultados melhores e mais estáveis que o modelo desenvolvido no projeto (R3D_18).

4.4 Melhores estratégias de postprocessing

O pós processamento atual melhorou o resultado em relação ao output *raw* do modelo. Este é feito a partir do output médio das probabilidades dada uma janela de tamanho N. Todavia, existem diversas outras possíveis estratégias como:

- **Filtros mais eficientes**: média ponderada ou filtros exponenciais que atribuem maior importância às previsões mais recentes. Além disso, a aplicação de técnicas baseadas em séries temporais, como Hidden Markov Models (HMM) ou Conditional Random Fields (CRF), pode contribuir para uma transição mais suave entre as classes e reduzir ruídos esporádicos na sequência de classificações.

- **Definir um overlap entre classificações**: atualmente, a classificação é feita em conjuntos de 16 frames sem overlap. Há a possibilidade de definir um overlap para possibilitar um melhor proveito da estratégia de pegar o output médio das probabilidades a partir de uma janela.
- **Penalidades para mudanças abruptas**: A inserção de penalidades para mudanças abruptas de classe em intervalos curtos — por meio de algoritmos de suavização ou penalização de transições rápidas — tende a favorecer a estabilidade da predição final.
- **Ensemble de pós-processamento**: Métodos de ensemble no pós-processamento como, por exemplo, a combinação de multiplas janelas de tamanho e overlap variados.