Relatório de Modelagem e Experimentação

Hugo Henrique Correia da Silva - 25 de maio de 2025

VISÃO GERAL

Neste estudo, foram avaliadas duas arquiteturas distintas de redes neurais para a tarefa conjunta de regressão de keypoints (pontos de articulação) e classificação de ações humanas a partir de imagens: uma rede convolucional tradicional, denominada PoseActionCNN, e uma arquitetura inspirada no modelo YOLO, chamada YOLOInspiredNet. A seguir, apresenta-se uma análise comparativa entre ambas as abordagens, considerando estrutura, diferenças funcionais e vantagens relativas.

OBJETIVOS

1. Desenvolva um pipeline de experimentação com pelo menos dois modelos distintos
2. Avalie o desempenho de cada abordagem com métricas apropriadas

PoseActionCNN

A PoseActionCNN adota uma estrutura convencional baseada em redes convolucionais profundas, composta por blocos de convolução (Conv2D), função de ativação não-linear (ReLU) e camadas de agrupamento máximo (MaxPooling). Após o processamento das imagens pelas camadas convolucionais, a rede é bifurcada em dois ramos independentes:

- Um ramo é responsável pela regressão dos keypoints, prevendo as coordenadas (x, y) de cada ponto por meio de camadas lineares.

- O outro ramo realiza a classificação da ação humana, utilizando camadas densas com ativação apropriada (como softmax ou logits).

Essa separação permite que cada tarefa seja tratada de forma especializada, promovendo simplicidade na estrutura e facilidade de treinamento.

YOLOInspiredNet

A `YOLOInspiredNet` baseia-se nos princípios do modelo YOLO (You Only Look Once), amplamente utilizado em tarefas de detecção em tempo real. Ao invés de bifurcar as tarefas, essa arquitetura realiza predições unificadas de coordenadas e classes diretamente a partir de uma única saída vetorial. A rede compartilha todo o aprendizado convolucional e produz uma saída combinada contendo:

- Coordenadas dos keypoints normalizadas (regressão).

- Classe da ação humana (classificação)

Dessa forma, a rede é otimizada para capturar simultaneamente as informações espaciais e semânticas, explorando de forma explícita a correlação entre a posição dos pontos e a ação realizada.

Comparação de Modelos

Para a prova de conceito, foi utilizado um subconjunto reduzido do dataset, contendo apenas 300 imagens por classe (total de 3 classes: *dancing*, *miscellaneous* e *sports*), totalizando 900 imagens para treinamento e validação. Este tamanho limitado de dados influencia diretamente no desempenho dos modelos.

Resultados Gerais

O modelo PoseActionCNN apresentou um F1-score macro de 0.3300 e uma acurácia de 0.3500. O erro quadrático médio (MSE) na regressão dos *keypoints* foi de 87.686,55.

Já o modelo YOLOInspiredNet obteve um F1-score macro inferior, de 0.1808, porém apresentou acurácia ligeiramente superior, 0.3722, e melhor desempenho na regressão dos keypoints, com MSE de 78.406,87.

Desempenho por Classe

No PoseActionCNN, a classe *miscellaneous* foi a única reconhecida com recall significativo (1.00), enquanto as classes *dancing* e *sports* não foram identificadas corretamente (f1-score 0.00). Isso indica um viés do modelo em favor da classe majoritária ou mais facilmente identificável.

O YOLOInspiredNet identificou bem a classe *dancing* (f1-score 0.54), mas falhou em reconhecer as classes *miscellaneous* e *sports* (f1-score 0.00 para ambas). Isso revela dificuldades na generalização para múltiplas classes, apesar do melhor desempenho na localização dos keypoints.

Considerações Finais

O desempenho limitado dos modelos está relacionado ao tamanho reduzido do dataset, com apenas 300 imagens por classe, o que dificulta o aprendizado robusto e a generalização para as diferentes posturas.

Recomenda-se a ampliação do conjunto de dados, além da aplicação de técnicas de balanceamento e aumento de dados (data augmentation), para melhorar a capacidade dos modelos em reconhecer todas as classes e refinar a predição dos *keypoints*.