UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MAURO GIOVANI BUCCO JUNIOR

PEDRO ANTONIO GUAPO DA ROCHA

TRABALHO DE VISÃO COMPUTACIONAL
POSE ESTIMATION

Trabalho para a Matéria de Visão Computacional com o desenvolvimento de uma rede neural convolucional que realiza pose estimation

Professor: Gustavo Benvenutti Borba

CURITIBA

2025

SUMÁRIO

1 INTRODUÇÃO

Na disciplina de Visão computacional com o professor Gustavo Benvenutti Borba, foi solicitado o desenvolvimento de uma rede neural convolucional capaz de determinar os pontos chaves (Keypoints) de uma imagem. A solução deveria utilizar como base uma das arquiteturas de redes estudadas em aula e incluir as etapas de pré-processamento das imagens do dataset, treinamento do modelo e validação dos resultados por meio de testes e métricas adequadas.

Para este trabalho, foi escolhida a rede MobileNetV2 como backbone, que se destaca por seu bom desempenho e baixo custo computacional. Sobre esse backbone, foram adicionadas camadas adicionais com operações de upsampling, com o objetivo de redimensionar os mapas de ativação para gerar heatmaps correspondentes às regiões de interesse. Esses mapas permitem a posterior identificação e visualização dos keypoints diretamente nas imagens originais.

A detecção de pontos-chave (Pose Estimation) é uma tarefa clássica da visão computacional que consiste em identificar a posição das juntas de uma pessoa. Com pontos como cabeça, ombros, cotovelos, punhos, quadris, joelhos e tornozelos, a partir de imagens ou vídeos. Cada ponto é representado por coordenadas (x, y) na imagem e, em conjunto, esses keypoints definem a pose.

Os arquivos utilizados no projeto estão disponíveis no github https://github.com/MauroBuccoJr/pose-estimation.

1.1 Objetivo Geral do Projeto

Desenvolver, implementar, testar e documentar uma rede neural convolucional capaz de processar imagens contendo seres humanos e estimar com precisão os pontos-chave (keypoints) do corpo. O modelo deve apresentar alta taxa de acerto na detecção das poses e baixo índice de falhas, atendendo aos requisitos de uma tarefa de visão computacional aplicada.

2 DESENVOLVIMENTO DA REDE NEURAL

2.1 Dataset Utilizado

O dataset utilizado neste trabalho é o Leeds Sports Pose (LSP), um conjunto de imagens utilizado para tarefas de estimativa de pose humana e detecção de pontos-chave. Ele contém 2.000 imagens coloridas de atletas em diversas posições corporais, com resolução média de 250 × 250 pixels. Cada imagem possui anotações com as coordenadas de 14 pontos-chave do corpo humano, incluindo cabeça, ombros, cotovelos, punhos, quadris, joelhos e tornozelos. Na Figura 1, temos uma representação do Dataset utilizado.

Figura 1 - Dataset LSP (LSP - Leed Sports Pose)

Fonte: Kaggle (2024)

2.2 Pré-Processamento

Os dados presentes nas anotações foram transpostos para o formato (2000, 14, 3), em que cada entrada representa os 14 keypoints da imagem, com suas respectivas coordenadas e a sua visibilidade na imagem. Para garantir uniformidade nas entradas da rede, todas as imagens foram redimensionadas para 256 × 256 pixels e os valores normalizados para o intervalo [0, 1].

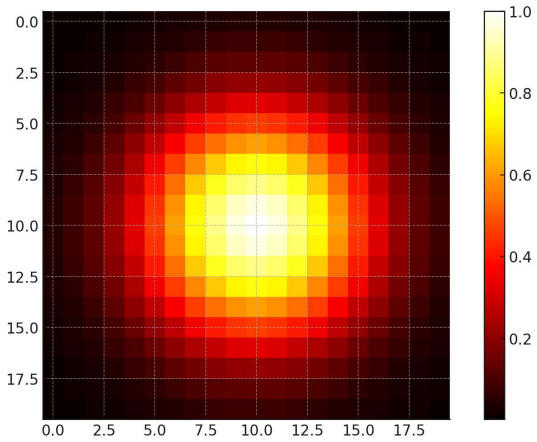
O pré-processamento também incluiu a transformação dos keypoints em mapas de calor (heatmaps) de dimensão 64 × 64 pixels. Para cada ponto, foi gerado um heatmap contendo uma marcação circular centrada na coordenada real do

keypoint. Esse processo permite que a rede aprenda a regressar regiões de probabilidade para os pontos, ao invés de coordenadas diretas.

2.3 Heatmaps com representação dos keypoints

Em vez de prever diretamente as coordenadas dos pontos-chave, uma técnica utilizada é a representação por heatmaps, como observado na figura 2. Nessa abordagem, para cada ponto-chave, é gerado um mapa bidimensional com a probabilidade da presença do ponto em cada pixel da imagem. A função que gera esse mapa retorna uma marcação circular gaussiana centrada no ponto escolhido.

Figura 2 - Exemplo de Heatmap



Fonte: PyTorch

Durante o treinamento, os keypoints anotados são transformados em heatmaps através dessa função. A rede neural é então treinada para regressar a esses heatmaps diretamente a partir da imagem de entrada, utilizando uma função de perda customizada entre os mapas previstos e os valores reais.

Essa abordagem apresenta diversas vantagens:

- A rede aprende uma distribuição espacial da posição esperada do ponto, em vez de apenas uma coordenada fixa.
- A saída da rede mantém a estrutura espacial, o que favorece a precisão dos pontos.
- É menos afetada pelas imprecisões nas anotações do dataset e pela ambiguidade de certos pontos corporais.

Após o processamento, a posição final de cada keypoint na imagem é obtida extraindo as coordenadas do ponto do heatmap que possui o valor mais alto através de outra função, e ajustado para o tamanho da imagem.

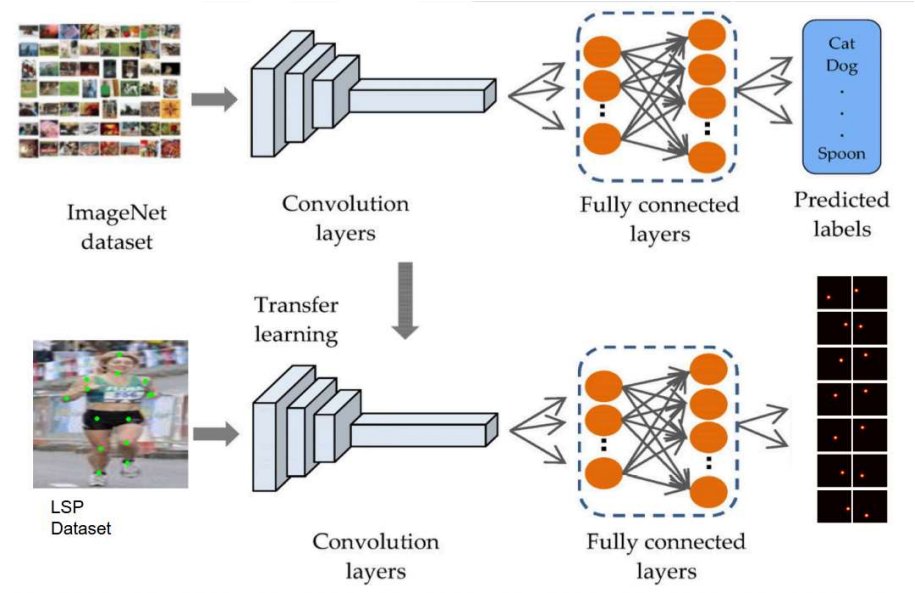
2.4 Arquitetura da Rede

A arquitetura proposta neste trabalho utiliza a MobileNetV2 como backbone para extração de características visuais. A rede pré-treinada no ImageNet foi utilizada sem as camadas de classificação (include_top=False), permitindo o uso de suas camadas convolucionais profundas como extrator de features.

Sobre a saída do backbone (de forma 8 × 8 × 1280), foram adicionadas camadas convolucionais com ativação ReLU e camadas de upsampling com fator 2, de modo a reconstruir heatmaps na resolução desejada (64 × 64). A camada final é uma convolução com 14 filtros e kernel 1 × 1, seguida de uma ativação sigmóide para limitar os valores entre 0 e 1.

A arquitetura completa, portanto, mapeia imagens de entrada com resolução 256 × 256 × 3 para uma saída de 64 × 64 × 14, onde cada canal representa o heatmap estimado de um ponto-chave. Na Figura 3, podemos observar o passo a evolução da arquitetura da rede.

Figura 3 - Arquitetura da Rede



Fonte: Autoria própria

2.5 Backbone MobileNetV2

O backbone de uma rede de pose estimation é o módulo responsável por extrair representações visuais profundas da imagem de entrada. Neste trabalho, utilizou-se a arquitetura MobileNetV2, uma rede convolucional leve e eficiente, especialmente projetada para dispositivos com restrições de recursos computacionais.

A MobileNetV2 é baseada em blocos lineares com conexões residuais invertidas e convoluções separáveis em profundidade (depthwise separable convolutions), o que reduz drasticamente o número de parâmetros e operações em comparação com arquiteturas mais pesadas como VGG ou ResNet. Apesar da redução de complexidade, a MobileNetV2 mantém um bom desempenho em tarefas de classificação e detecção.

Ao utilizar MobileNetV2 como backbone, foram removidas as camadas de classificação originais (top layers) e adicionadas camadas adicionais de convolução e upsampling (UpSampling2D) para aumentar a resolução espacial da saída. Isso possibilita a geração de heatmaps com dimensão 64×64, permitindo a regressão precisa dos keypoints posteriormente.

A utilização de um backbone pré-treinado no ImageNet contribui para uma convergência mais rápida do modelo e melhora a generalização, o que é uma vantagem quando o tamanho do dataset é limitado como no caso do LSP.

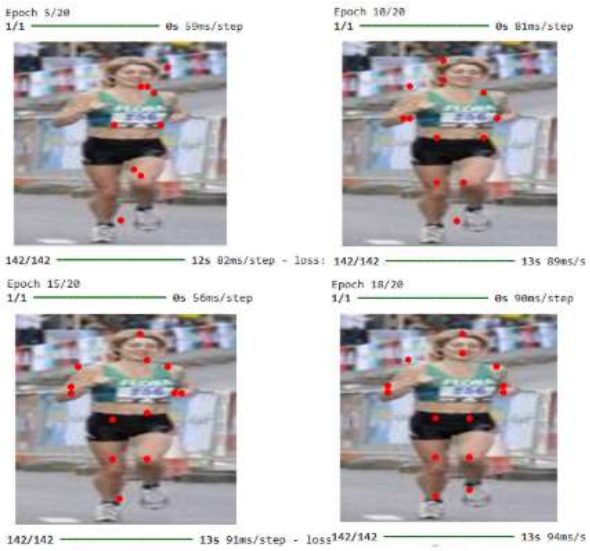
2.6 Treinamento do Modelo

O modelo foi compilado com o otimizador Adam, utilizando taxa de aprendizado inicial de 1×10^{-4}. A função de perda escolhida foi customizada, combinando o erro quadrático médio (MSE) entre os pixels dos heatmaps preditos e os heatmaps do ground truth, com o erro médio quadrático entre as coordenadas do centro dos heatmaps preditos com as coordenadas verdadeiras dos keypoints.

Durante o treinamento, foram utilizados batches de 12 imagens e 15 épocas de treinamento. O dataset foi dividido em três partes, 85% para treinamento, 10% para validação e 5% para teste.

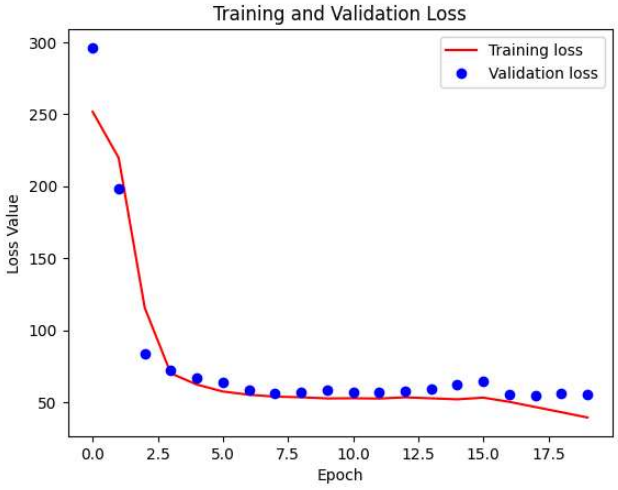
Como métrica qualitativa, uma função de callback foi adicionada com o propósito de plotar a projeção dos keypoints estimados sobre as imagens originais a cada final de época de treinamento. A seleção da melhor época foi feita com base na visualização dos resultados qualitativos e observação do comportamento do loss ao longo das épocas, uma vez que não foi utilizada uma métrica quantitativa formal como PCK. As Figuras 4 e Figura 5 apresentam a evolução do treinamento e função de perda durante o treinamento, mostrando que o modelo convergiu de forma estável ao longo das épocas, tanto no conjunto de treino quanto no de validação.

Figura 4 - Resultados da Rede



Fonte: Autoria própria

Figura 5 - Resultados da Rede



Fonte: Autoria própria

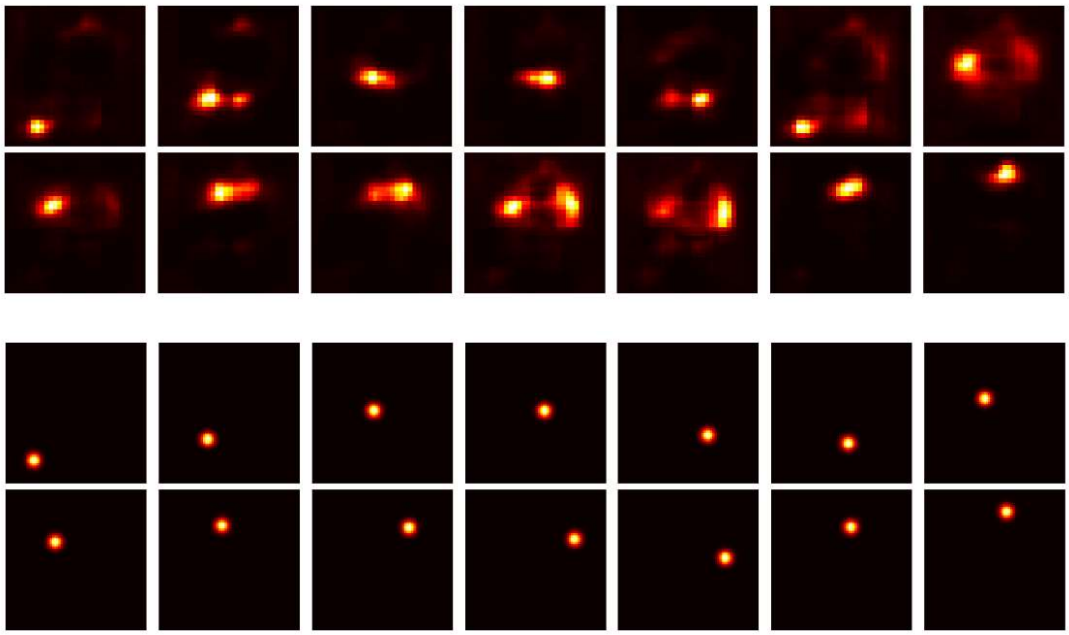
A implementação foi realizada com TensorFlow 2.x e Keras, utilizando também bibliotecas auxiliares como NumPy, OpenCV e Matplotlib para o pré-processamento, visualização e análise dos resultados. O ambiente de desenvolvimento utilizado foi o Google Colab com aceleradores de hardware (GPU).

3 RESULTADOS

3.1 Heatmaps previstos vs Ground Truth

Para avaliar o desempenho do modelo, foi realizada a comparação visual entre os Heatmaps obtidos pela previsão do modelo e os heatmaps do ground truth. Na Figura 6 abaixo, temos uma imagem representando essa comparação.

Figura 6 - Heatmaps previstos vs Ground Truth



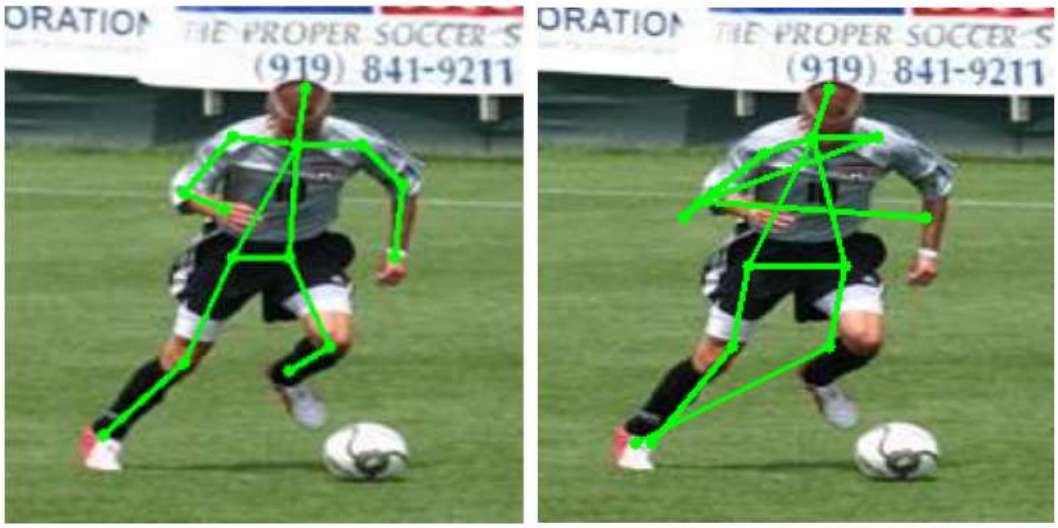
Fonte: Autoria própria

3.2 Imagem com pontos-chave estimados

A Figura 7 apresenta os resultados da projeção dos pontos-chave estimados pela rede diretamente sobre a imagem de entrada. Para cada heatmap de saída, foi obtida a coordenada do pixel de maior ativação, considerada como a posição estimada do keypoint correspondente. Posteriormente, esses pontos foram

conectados com linhas e posicionados sobre as imagens de acordo com a estrutura esquelética do corpo humano, facilitando a interpretação da pose prevista.

Figura 7 - Comparação ground truth vs resultados da rede



Fonte: Autoria própria

4 DISCUSSÕES QUALITATIVA E LIMITAÇÕES

Analisando-se qualitativamente as predições realizadas pelo modelo treinado, observa-se que a precisão na localização dos keypoints ainda não está adequada. Isso se deve, em grande parte, ao fato de os heatmaps gerados serem amplos e difusos, com regiões de ativação pouco concentradas ao redor dos keypoints, conforme ilustrado nas Figuras 4 e 5. Tais resultados indicam que a rede não foi capaz de aprender representações suficientemente boas para localizar os keypoints com exatidão.

Essas limitações estão associadas à natureza restrita do conjunto de dados utilizado. O dataset LSP contém apenas 2.000 imagens, com variação limitada de poses, roupas, fundos e condições de iluminação, e não contempla múltiplas pessoas por imagem ou movimentos altamente não convencionais.

Uma estratégia potencial para mitigar esse problema consiste na aplicação de técnicas de aumento de dados (data augmentation), como rotações, espelhamentos horizontais, escalonamentos e alterações de brilho. Essa abordagem permite ampliar artificialmente a diversidade do conjunto de treinamento, o que pode

contribuir para melhorar a capacidade de generalização do modelo e, consequentemente, refinar a precisão na detecção dos keypoints.

Outra possibilidade de aprimoramento do modelo seria a ampliação da profundidade da arquitetura, por meio da adição de camadas convolucionais adicionais. Esse ajuste poderia aumentar a capacidade da rede de extrair características mais complexas e representações mais refinadas dos dados de entrada, o que é especialmente útil em tarefas com alto grau de variação espacial, como a estimativa de pose.

Além disso, pode-se considerar a aplicação de técnicas de fine-tuning, em que os pesos da rede pré-treinada são parcialmente ajustados durante o treinamento supervisionado. Essa abordagem é particularmente recomendada em casos em que o modelo apresenta estagnação na função de perda ao longo das épocas, indicando limitação na capacidade de adaptação à tarefa específica.