Universidade Fernando Pessoa

Mestrado em Computação Móvel

Inteligência Artificial

Rodrigo Soares - 36824 e Luísa Costa - 37151

Projeto Prático de Inteligência Artificial

Reconhecimento de actividade humana (human activity recognition) usando deep learning (ML) e vídeo datasets

---

Porto

Junho/2021

Universidade Fernando Pessoa

Praça 9 de Abril, 349

P-4249-004 Porto

Tel. +351-22550.82.70

Fax. +351-22550.82.69

geral@ufp.pt

1. Introdução

O reconhecimento de atividades humanas com base em vídeo é uma área de estudo que tem vindo a crescer, no entanto, continua a ser uma área complexa devido às variadas nuances existentes, tais como, posição da câmera, luminosidade e escala da cena no processo de captura além das possíveis combinações de ações que os actores podem efetuar para uma mesma actividade[1].
Este relatório terá como objectivo analisar uma implementação de um projeto de classificação de vídeo de atividade humana[2] e posteriormente fazer uma análise comparativa entre diferentes modelos de redes neuronais com o objectivo de determinar qualitativamente quais as mais adequadas para a realização da tarefa.

2. Descrição do problema

Em essência, para se poder fazer a classificação de vídeos de atividades humanas, antes é preciso compreender o que é o reconhecimento de atividades humanas. A diferença de esta classificação para outra, é que na classificação humana, uma análise estática de uma atividade não nos permite determinar qual é essa atividade, sendo então necessário múltiplos pontos temporais para a determinar. No trabalho explorado[2], o autor faz esta comparação com o exemplo de uma pessoa a fazer um salto. Não é possível determinar através de uma só imagem se a pessoa se encontra no meio de um salto ou se numa queda livre, contudo com acesso ao vídeo completo, já se pode classificar.

3. Estado da Arte

A área de reconhecimento de atividade humana tem sido bastante procurada nos últimos anos devido à sua versatilidade. Existem inúmeras maneiras de registar atividade humana, tais como usando sensores de movimento, sons e, como utilizado neste trabalho, capturas de vídeo. Existem, no entanto, desafios relacionados com a previsão da atividade humana com base em vídeos tais como determinar quais os parâmetros e dimensões necessárias, sendo estes fatores altamente dependentes dos recursos utilizados.
Nos dias atuais temos uma panóplia de variantes das redes neuronais convolucionais, como por exemplo as *Time information fusion CNN's* (algoritmos que fundem informação ao longo do domínio do tempo), *single-frame CNN* (abordagem utilizada para perceber a precisão da classificação), *slow fusion*, *late fusion* e *early fusion* (abordagens semelhantes, apenas com fusão de frames em momentos diferentes).
Mais recentemente foram implementadas CNN's com LSTM's (composto por células, em que cada célula pode processar sequencialmente os dados).

4. Descrição do trabalho realizado

Numa fase inicial, o objectivo foi de implementar o projecto já existente[2], estudar a sua estrutura e verificar a sua funcionalidade. Esse projeto tem como base o dataset UCF50[3] o qual é composto por cinquenta classes de atividades humanas tais como salto à corda, corridas de cavalos, tocar violino, entre outras. O autor inicialmente analisa o problema de entender o reconhecimento de atividade humana e a sua relação com a classificação de vídeos, tendo por base o contexto entre as diferentes frames de um vídeo, permitindo assim distinguir entre diferentes atividades via uma análise temporal. Uma possível abordagem para solucionar este problema, como o autor indica, seria fazer uma média ponderada dos resultados de n frames seguidas para

classificar esse segmento de vídeo, abordagem esta aplicada no projeto. Contudo esta solução não é um método de classificação de vídeo com base em redes neuronais, mas sim um melhoramento à análise de frames individuais.
O autor também faz menção deste aspecto e indica oito métodos diferentes de classificação de vídeo com base em redes neuronais, sendo estes:

- *Single-Frame CNN*:

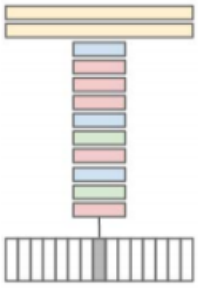


Fig 1 - Estrutura de uma Single-Frame CNN

Abordagem utilizada para perceber a precisão da classificação passando frames individuais para uma CNN.
CNN são redes neuronais convolucionais que se tornaram dominantes em múltiplas tarefas de visão computacional pois "aprendem" automática e adaptativamente por meio de retropropagação, usando vários blocos de construção, como camadas de convolução, camadas de pool e camadas totalmente conectadas.

- *Late Fusion*

Este modelo como indicado na figura 2 separa duas redes até à última camada convolucional com uma distância de N frames e posteriormente funde os dois fluxos numa primeira camada totalmente conectada, tendo assim a possibilidade de comparar ambas as saídas.

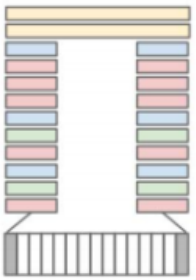


Fig 2 - Estrutura de uma Late Fusion

- *Early Fusion*

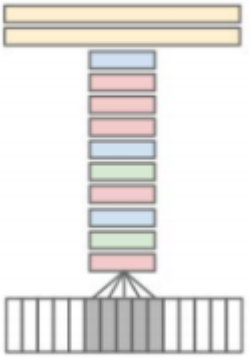


Fig 3 - Estrutura de uma Early Fusion

Este método por sua vez combina múltiplas entradas, aplicando filtros na primeira camada convolucional.

- *CNN com LSTM's*

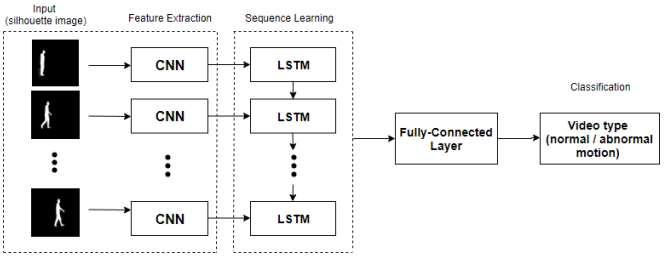


Fig 4 - Estrutura de uma CNN com LSTM's

Este algoritmo, como podemos ver na figura 4, usa as CNN para extrair as características de cada frame e este output é utilizado pela rede LSTM que o irá fundir temporalmente.

- *Deteção de pose e LSTM*

Usa-se o modelo de detecção de pose para obter os pontos-chave de um

corpo para cada frame do vídeo e, em seguida, usa-se esses pontos-chave numa rede LSTM para determinar a atividade que está a ser realizada.

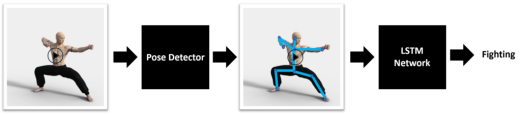


Fig 5 - Estrutura de modelo usando detecção de posse e LSTM

- *Optical Flow e CNN's*

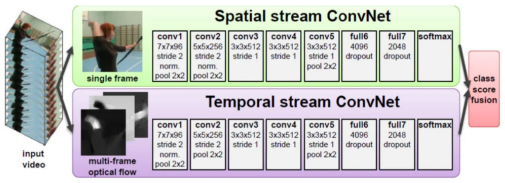


Fig 6 - Estrutura de modelo usando Optical Flow e CNN's

Nesta abordagem, são usados dois fluxos paralelos de redes convolucionais: a stream que é conhecida como Spatial Stream (seleciona uma única frame do vídeo e executa várias CNN kernels com base nas suas informações espaciais fazendo posteriormente uma previsão) e o fluxo na parte inferior, chamado de fluxo temporal (seleciona os fluxos ópticos de cada frame adjacente após ser aplicada a técnica early fusion e, em seguida, faz merge com a informação proveniente do vídeo para fazer a previsão).
No final, a média entre ambas as probabilidades previstas é realizada para obter as probabilidades finais.

- *SlowFast Networks*

Semelhante ao método anterior, esta abordagem também possui dois fluxos paralelos: fluxo no topo (slow branch) que trabalha com fluxos de baixa taxa de frames e possui muitos canais em cada camada para o processamento detalhado de cada frame e o fluxo na parte inferior (fast branch) que tem poucos canais mas trabalha com fluxos de alta taxa de frames.

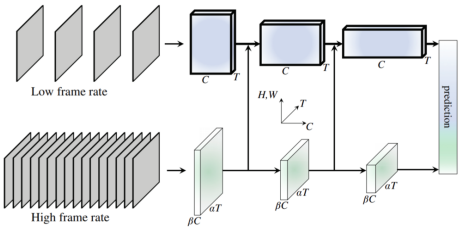


Fig 7 - Estrutura de modelo usando redes SlowFast

- *3D CNN's / Slow Fusion*

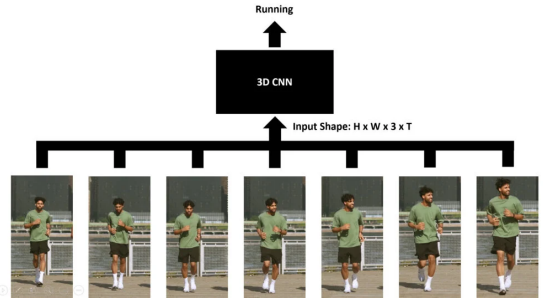


Fig 8 - Estrutura de modelo usando CNN's 3D / Slow Fusion

Esta abordagem usa uma rede de convolução 3D que permite o processamento de informações temporais e espaciais usando uma CNN tridimensional. Ao contrário do late fusion e do early fusion, este método funde as informações temporais e espaciais lentamente em cada camada da rede da CNN.

Destes métodos o autor vai implementar o denominado de *Single-Frame CNN* através da framework Keras e para tal, analisou a estrutura do dataset, determinou o tamanho de cada frame (64 por 64 pixels), selecionou quatro classes para classificação (Taichi, HorseRace, Swing e WalkingWithDog) e dividiu o dataset em grupos de teste e de treino num rácio de 80% e 20% respetivamente. O modelo

CNN implementado é composto por duas camadas CNN seguidas de algumas camadas de normalização e de pooling como se pode observar na figura 9.

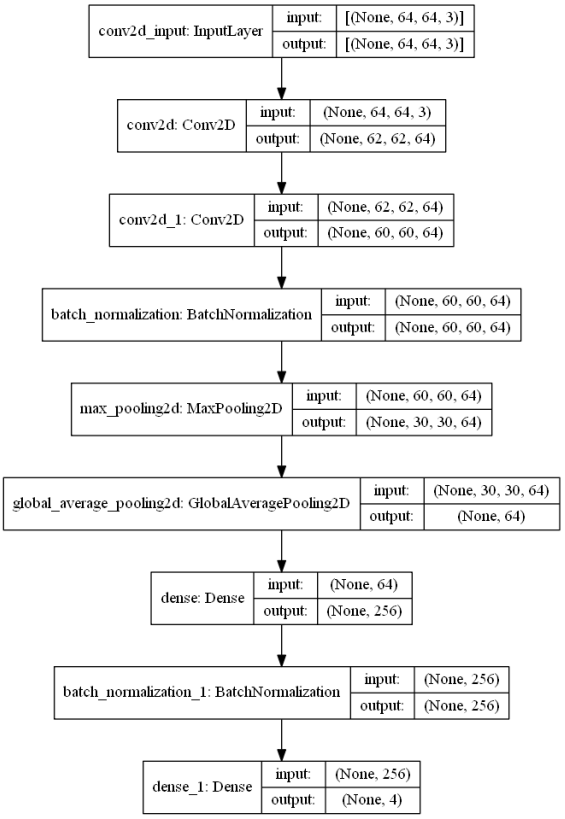


Fig 9 - Estrutura do modelo criado usando single-frame CNN

Numa segunda fase foi também implementado para este relatório o método CNN com LSTM's o qual é composto por duas camadas CNN encapsuladas numa distribuição temporal (necessária para a implementação com LSTM) e de camadas de dropout e de pooling também encapsuladas, seguidas da camada LSTM como se pode observar na figura 10.

Numa fase final efetuou-se previsões com vídeos para não só validar o funcionamento dos modelos, como também para verificar a sua precisão de classificação das atividades humanas presentes, mesmo que cada vídeo contenha múltiplas atividades em diferentes períodos.

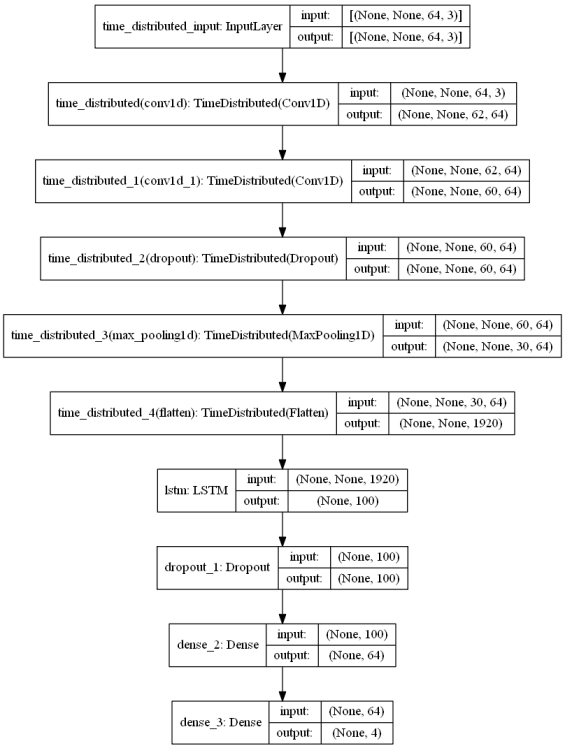


Fig 10 - Estrutura do modelo criado usando CNN com LSTM

5. Análise de Resultados

5.1 Histórico de Avaliação dos Modelos

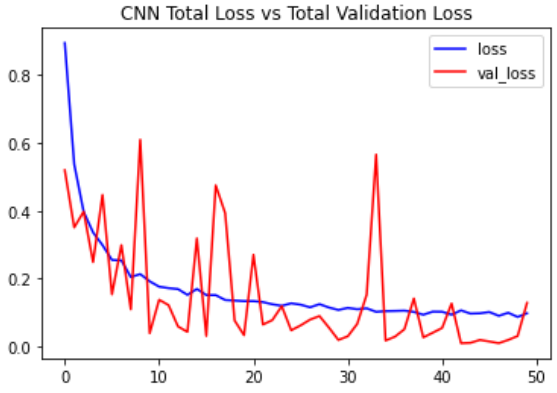
	CNN	CNN-LSTM
Loss	0.1565	0.0331
Accuracy	0.9675	0.9931

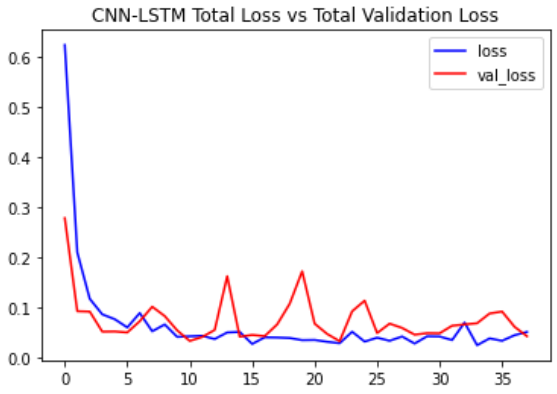
Como podemos ver na tabela acima descrita, ao avaliar a precisão dos dois modelos com base no dataset de teste, pode-se aferir que o modelo CNN-LSTM teve uma melhoria tanto em termos de precisão como em perda.

5.2 Métricas dos Modelos

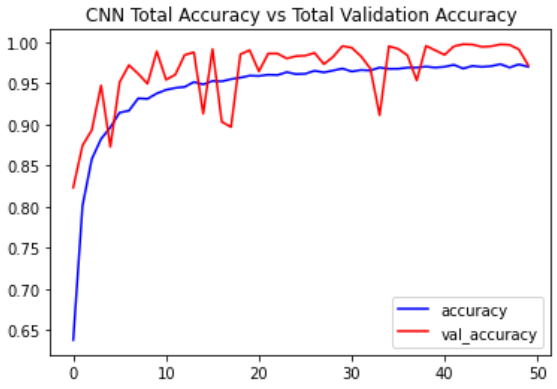
Os gráficos que vamos observar em seguida são gráficos que foram construídos sobre o domínio do tempo, ou seja, foram construídos durante o treino do dataset.

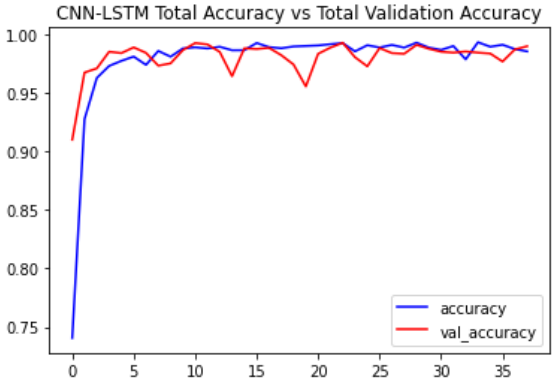
Relativamente às métricas dos modelos, comecemos por observar os gráficos relativos à perda.





Como podemos inferir, os valores da perda no método CNN-LSTM são bastante mais rápidos a convergir para valores baixos do que no CNN. Também se pode verificar que a validação da perda é mais estável no modelo CNN-LSTM.
Para terminar a avaliação das métricas do modelo, pode-se observar os gráficos relativos à precisão.





O mesmo resultado pode ser observado nos gráficos, neste caso para valores crescentes, nos quais ao longo do tempo vai havendo um aumento logarítmico da precisão do dataset.

5.3 Previsões em Vídeos:

Posteriormente foi feita uma análise preditiva com base numa média de classificação de uma janela de n frames de um vídeo de teste. Este vídeo é composto por quatro segmentos, cada um referente a uma das classes dos modelos treinados.
Para se comparar a eficácia dos modelos, foram feitos quatro testes, ora para cada modelo, ora para dois tamanhos de janela (janela de tamanho um e de tamanho vinte e cinco).

5.3.1 Tamanho de janela 1 (frame a frame)

O modelo CNN aparentou ser relativamente estável, com dificuldades em cenas onde se perde o contexto da atividade, contudo apresentou bastante confusão num cenário onde deveria ser a classe Taichi a qual ficava a oscilar com a classe WalkingWithDog.
Por sua vez, o modelo CNN-LSTM aparentou ser bastante estável, todavia apresentou bastante confusão num cenário onde deveria ser classe WalkingWithDog a qual ficava a oscilar com a classe Swing. É de notar que a qualidade do vídeo nesse

segmento é menor que o adequado, o que pode ter acrescido esta dificuldade.

5.3.1 Tamanho de janela 25

O modelo CNN aparentou ser bastante estável, contudo confuso num cenário onde deveria ser a classe Taichi a qual fica a oscilar com a classe WalkingWithDog. Esta oscilação é idêntica mas menos frequente à presente no tamanho de janela igual a um.
Por sua vez, o modelo CNN-LSTM aparentou ser bastante estável, porém apresentou bastante confusão num cenário onde deveria ser classe WalkingWithDog a qual ficava a oscilar com a classe Swing. Esta oscilação é idêntica mas menos frequente à presente no tamanho de janela igual a um.

5.4 Previsões com Média em Vídeos:

Os valores abaixo descritos são valores relativos à previsão, ou seja, são o output de uma função que retira n frames (neste caso vinte e cinco frames) do vídeo a ser previsto, indo classificar frame a frame com o respectivo modelo e no final é realizada a média dessas previsões. Este método não é adequado para situações onde se deseje sequências de frames como é o caso do modelo CNN-LSTM.

Vídeo 1 - Taichi

	CNN (%)	CNN-LSTM (%)
TaiChi	**~100**	**~100**
HorseRace	0.2	0.00045
WalkingWithDog	0.059	0.0091
Swing	0.0097	0.016

Vídeo 2 - Swing

	CNN (%)	CNN-LSTM (%)
Swing	**~94**	**~82**
WalkingWithDog	5.5	~18
HorseRace	~0	0.05
TaiChi	~0	0.012

Vídeo 3 - WalkingWithDog

	CNN (%)	CNN-LSTM (%)
WalkingWithDog	**~91**	~19
Swing	8.9	**~81**
TaiChi	~0	0.0025
HorseRace	~0	0.017

Vídeo 4 - HorseRace

	CNN (%)	CNN-LSTM (%)
HorseRace	~49	~69
Swing	**~51**	~29
WalkingWithDog	0.2	2.5
TaiChi	0.0086	0.0095

6. Conclusões

Como se pode observar, os resultados nos vídeos um e dois são bastante desejáveis, porém no vídeo três pode-se observar que o modelo CNN-LSTM teve dificuldades na classificação. Isto poderá ser devido ao facto de o modelo estar mais adequado

para uma análise de frames em sequência e não individualmente, além de o próprio modelo possivelmente necessitar de mais elementos de treino pois já na fase de previsões em vídeos, existia uma constante confusão entre as classes WalkingWithDog e Swing. No vídeo quatro pode-se observar pouca certeza no modelo CNN em relação à classe a escolher, certeza esta melhorada no modelo CNN-LSTM.
Isto leva a concluir que, num cenário onde se tenha acesso a sequências de frames de vídeos de uma atividade a classificar será mais adequado utilizar o modelo CNN-LSTM ao invés do CNN de frame singular, ou possivelmente um outro modelo com análise ao longo do tempo.
Possíveis melhorias que poderiam ser implementadas iriam incidir principalmente ora na implementação de outros modelos para comparação, ora na modificação da estrutura dos modelos implementados para tentar obter melhores resultados, pois, como é o caso no modelo CNN-LSTM, este foi concebido com recurso a camadas Conv1D, camadas essas não sendo as mais adequadas para processamento e classificação de elementos visuais (imagem/vídeo).

7. Referências

1. A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," in IEEE Access, vol. 6, pp. 1155-1166, 2018, doi: 10.1109/ACCESS.2017.2778011.

2. Learn OpenCV. (2021, May 3). Introduction to Video Classification and Human Activity Recognition. Learn OpenCV | OpenCV, PyTorch, Keras, Tensorflow Examples and Tutorials. https://learnopencv.com/introduction-to-video-classification-and-human-activity-recognition/

3. Kishore K. Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. Mach. Vision Appl. 24, 5 (July 2013), 971–981. DOI:https://doi.org/10.1007/s00138-012-0450-4

Anexos

- Github do Projecto: https://github.com/Infor-Master/IART_Projecto
- Dataset: https://www.crcv.ucf.edu/data/UCF50.rar