Reconhecimento de ações humanas em vídeos por meio de transferência de aprendizado

Lucas Messias Valério
lucasvalerio@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil.

Priscila Tiemi Maeda Saito
psaito@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil.

RESUMO

OBJETIVO: Este projeto tem como objetivo o estudo, avaliação e proposta de técnicas para aprendizado e reconhecimento de ações humanas em vídeos. MÉTODOS: A metodologia de desenvolvimento proposta considera técnicas de Aprendizado Profundo (*Deep Learning*). Inicialmente, foi utilizado o modelo Inception V3, uma Rede Neural Convolucional já treinada e em seguida aplicada a estratégia de Transferência de Aprendizado (*Transfer Learning*), a qual possibilita o retreinamento da rede para as diferentes bases de dados. Para os experimentos foram utilizadas as bases de dados Weizmann e KTH, as quais são públicas e comumente utilizadas por diversos trabalhos na literatura. A metodologia apresentada considera uma análise do vídeo frame a frame, não levando em consideração dados temporais. Dessa forma, é também apresentada uma possível solução para esse problema. RESULTADOS: Após a aplicação da metodologia proposta foram obtidos os resultados de acurácia final de 89,4% para a base Weizmann e 74% para a base KTH. CONCLUSÕES: A utilização de estratégias de Transferência de Aprendizado apresenta um resultado satisfatório levando em consideração que a rede, a princípio genérica e em seguida retreinada para reconhecer ações específicas de cada banco de dados, apresenta vantagens em relação a recursos e tempos computacionais, além dos valores de acurácias de classificação obtidos.

PALAVRAS-CHAVE: Aprendizado profundo. Transferência de aprendizado. Ações humanas. Rede neural convolucional.

INTRODUÇÃO

Com o passar dos anos, o avanço da tecnologia digital tem permitido mais facilidades como gravação de dados (vídeos, imagens, áudio, entre outros) em alta qualidade e grandes capacidades de armazenamento e processamento desses dados. No entanto, muitas dessas informações geradas apenas são armazenadas e não passam por nenhum tipo de análise e/ou organização, seja por máquina ou por um operador, para posterior recuperação de informações.

A área de aprendizado de máquina vem alterando significativamente como os computadores processam grandes quantidades de dados. Técnicas de reconhecimento e aprendizado de padrões específicos de um domínio possibilitam a análise e classificação automática dos dados para diversas aplicações. Por exemplo, estudos e desenvolvimento de abordagens considerando o domínio de reconhecimento de ações humanas tem beneficiado inúmeras aplicações, como: vigilância em geral (LIU; GU; KAMIJO, 2016), casas inteligentes (CHOI; KIM; OH, 2013), monitoramento de saúde (WOLF, 2014), realidade aumentada (MAQUEDA, 2015), entre outras.

A área de reconhecimento de ações humanas é vasta, e existem diversas implementações de sistemas robustos na literatura. Além disso, devido à grande disponibilidade de conjuntos de vídeos, bem como restrições de tempo e recursos computacionais apresentadas por determinadas aplicações, torna-se essencial o desenvolvimento de técnicas mais eficazes e eficientes. Neste contexto abordagens baseadas em aprendizado profundo têm sido amplamente exploradas e aperfeiçoadas (JI, 2013).

MÉTODOS

Primeiramente é necessário realizar o tratamento dos conjuntos de vídeos. A base Weizmann apresenta 10 classes de ações e 9 atores, totalizando 90 vídeos em formato .avi, os quais quando convertidos para .jpg, geraram 5.760 arquivos. A base KTH possui 6 classes de ações e 25 atores, sendo que cada um deles possui 4 vídeos, totalizando 600 vídeos. Devido ao hardware limitado utilizado nos experimentos (Processador Core i5 6400, 8GB de RAM e Placa de Vídeo GTX 1060), não foi possível trabalhar com toda essa quantidade de vídeos. Dessa forma, foram selecionados 2 vídeos por atores, gerando um total de 300 vídeos, que quando convertidos para imagens originaram 153.736 arquivos.

Após a preparação dos dados de entrada, foi implementado, utilizando o Framework TensorFlow, um algoritmo em Python responsável por realizar a Transferência de Aprendizado (*Transfer Learning*), a qual retreina uma parte de um modelo de aprendizagem de máquina pré-existente para uma nova finalidade.

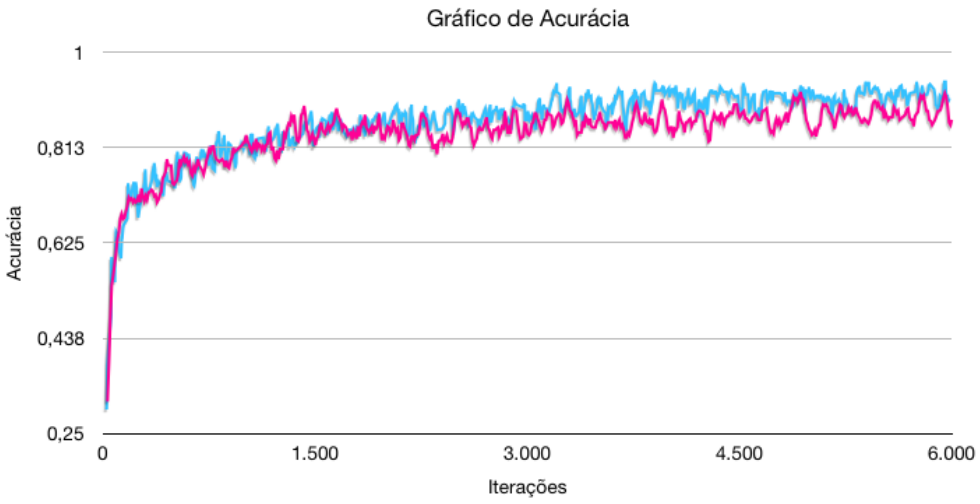
O modelo utilizado como base foi o Inception V3 (SZEGEDY, 2016), uma rede neural já existente desenvolvida pelo Google para disputar a competição do ImageNet (RUSSAKOVSKY, 2015). No presente trabalho, especificamente é considerada a camada de Estrangulamento (*Bottleneck*) do Inception V3, a qual atua antes da camada final de *Softmax*. Em seguida, apenas uma camada de toda essa rede precisa ser retreinada, já que as camadas anteriores já aprenderam funções úteis e generalizáveis, como por exemplo, a detecção de bordas.

Para validação dos conjuntos, o algoritmo se encarrega de dividir a entrada de dados em 3 partes: Treinamento, Validação e Teste, definindo 80% do total de imagens de entrada para treinamento, 10% para executar validações frequentes durante o treinamento e os 10% restantes para testes para verificar o desempenho do classificador.

RESULTADOS

Considerando a base Weizmann, com 5.760 arquivos, o tempo de treinamento durou cerca de 20 minutos com 4.000 iterações. Após essas iterações, as acurácias mantiveram-se estáveis, sendo obtida uma acurácia final de 89,4%, como pode ser visto na Figura 1.

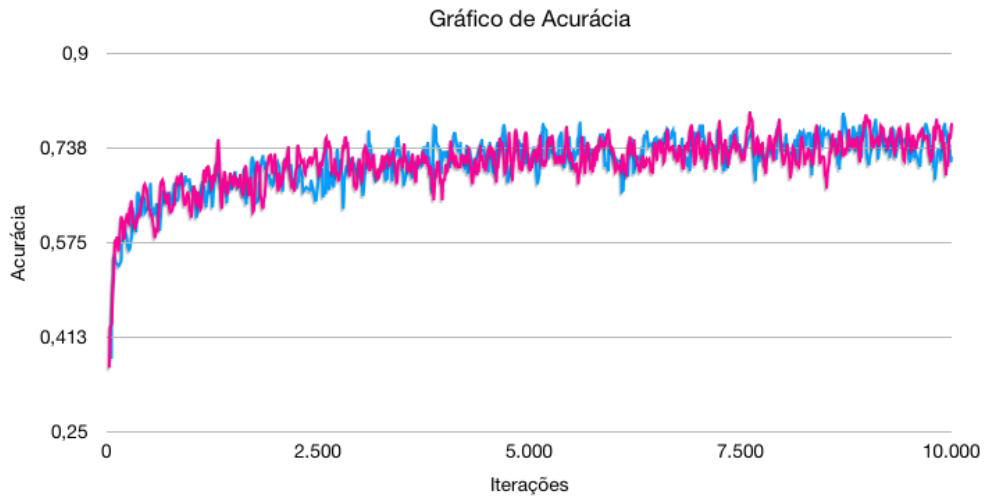
Figura 1 – Acurácias utilizando a base Weizmann (BLANK, 2005)



Fonte: Autoria própria (2017).

Para a base KTH com 153.736 arquivos, o tempo de treinamento durou cerca de 8 horas com 10.000 iterações. Nesse caso, mais iterações foram necessárias para obtenção de melhores resultados, sendo obtida uma acurácia final de 74% (Figura 2).

Figura 2 – Acurácias utilizando a base KTH (SCHULDT; LAPTEV; CAPUTO, 2004)



Fonte: Autoria própria (2017).

DISCUSSÃO

Nas Figuras 1 e 2, a cor azul representa o resultado obtido com a validação no treinamento, e a cor rosa o resultado com a avaliação no conjunto de testes. Para a base Weizmann (Figura 1), pode-se observar claramente que para a etapa de validação no treinamento, os resultados foram melhores em relação à etapa de validação no teste. Isso pode ser explicado devido ao fato de que a rede aprendeu características mais especificas da imagem e não tão gerais. Tal fato pode ser um problema, por exemplo, da base de dados não estar muito bem tratada, visto que, em alguns vídeos os atores saem de cena quando realizam alguma ação, e a câmera acaba por filmar um fundo estático.

Analisando a base KTH (Figura 2) é possível observar que embora as acurácias não sejam tão elevadas, as validações no treinamento e no teste apresentaram resultados equivalentes, demonstrando que a rede aprendeu as características generalizadas das imagens. Além dos resultados de acurácias, outras medidas foram avaliadas, as quais não foram apresentadas no presente trabalho devido às limitações de espaço, mas que podem ser acessadas em (VALÉRIO, 2017).

CONSIDERAÇÕES FINAIS

Pode-se concluir que os resultados obtidos foram satisfatórios, partindo do princípio que a rede era genérica e foi retreinada para identificar casos específicos de dados. A metodologia proposta possibilitou atingir elevados níveis de acurácias, bem como otimizar a utilização dos recursos e tempos computacionais.

Um dos maiores problemas dessa metodologia é a análise de vídeos frame a frame, a qual não leva em consideração dados temporais. Como trabalho futuro, propõe-se o tratamento dos dados de entradas não como frames discretos, mas sim como sequências de fato, considerando os dados temporais, por meio da implementação de uma Rede Neural Recorrente ligada à saída da Rede Neural Convolucional obtida até o momento. Com o uso da Rede Neural Recorrente, além de processar um frame, seria possível guardar informações dos frames anteriores, realizando assim um aprendizado de todo o movimento e não só a classificação de uma imagem estática.

Human action recognition in videos through transfer learning

ABSTRACT

OBJECTIVE: This project aims to study, evaluate and propose techniques for learning and recognizing human actions in video. **METHODS**: The proposed development methodology considers Deep Learning techniques. Initially, the Inception V3 model was used, a Convolutional Neural Network already trained and then applied to the Transfer Learning strategy, which allows the re-training of the network for the different databases. For the experiments we used the Weizmann and KTH databases, which are public and commonly used by several works in the literature. The presented methodology considers a analysis of the video frame by frame, not taking into account temporal data. In this way, a possible solution to this problem is also presented. **RESULTS:** After the application of the proposed methodology, the results of final accuracy of 89.4% for the Weizmann base and 74% for the KTH base. **CONCLUSIONS:** The use of Learning Transfer strategies presents a satisfactory result taking into account that the network, generally generic and then retrained to recognize specific actions of each database, presents advantages in relation to resources and computational times, besides the values of Obtained.

KEYWORDS: Deep learning. Transfer learning. Human actions. Convolutional neural network.

AGRADECIMENTOS

À acurada orientação da Profa. Dra. Priscila Tiemi Maeda Saito e a Universidade Tecnológica Federal do Paraná (UTFPR) pela concessão de uma bolsa para realização desta pesquisa por meio do Programa de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação – PIBITI 2016/2017.

REFERÊNCIAS

LIU, JINGWEN; GU, YANLEI; KAMIJO, SHUNSUKE. JOINT CUSTOMER POSE AND ORIENTATION ESTIMATION USING DEEP NEURAL NETWORK FROM SURVEILLANCE CAMERA. IN: **ISM**, 2016. P. 216-221.

CHOI, SUNGJOON; KIM, EUNWOO; OH, SONGHWAI. HUMAN BEHAVIOR PREDICTION FOR SMART HOMES USING DEEP LEARNING. IN: **RO-MAN,** 2013. P. 173-179.

WOLF, CHRISTIAN ET AL. **EVALUATION OF VIDEO ACTIVITY LOCALIZATIONS INTEGRATING QUALITY AND QUANTITY MEASUREMENTS.** COMPUTER VISION AND IMAGE UNDERSTANDING, V. 127, P. 14-30, 2014.

MAQUEDA, ANA I. ET AL. HUMAN-ACTION RECOGNITION MODULE FOR THE NEW GENERATION OF AUGMENTED REALITY APPLICATIONS. IN: **ISCE,** 2015. P. 1-2.

JI, SHUIWANG ET AL. 3D CONVOLUTIONAL NEURAL NETWORKS FOR HUMAN ACTION RECOGNITION. **TPAMI,** V. 35, N. 1, P. 221-231, 2013.

SZEGEDY, CHRISTIAN ET AL. RETHINKING THE INCEPTION ARCHITECTURE FOR COMPUTER VISION. IN: **CVPR.** 2016. P. 2818-2826.

RUSSAKOVSKY, OLGA ET AL. **IMAGENET LARGE SCALE VISUAL RECOGNITION CHALLENGE.** IJCV, V. 115, N. 3, P. 211-252, 2015.

BLANK, MOSHE ET AL. ACTIONS AS SPACE-TIME SHAPES. IN: **ICCV,** 2005. P. 1395-1402.

SCHULDT, CHRISTIAN; LAPTEV, IVAN; CAPUTO, BARBARA. RECOGNIZING HUMAN ACTIONS: A LOCAL SVM APPROACH. IN: **ICPR,** 2004. P. 32-36.

VALÉRIO, L. M. **RECONHECIMENTO DE AÇÕES HUMANAS EM VÍDEOS UTILIZANDO APRENDIZADO PROFUNDO.** UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ. 2017.

Recebido: 31 ago. 2016.
Aprovado: 02 out. 2016.
Como citar:
VALÉRIO, L. M.; SAITO, P. T. M. Reconhecimento de ações humanas em vídeos por meio de transferência de aprendizado. In: SEMINÁRIO DE INICIAÇÃO CIENTÍFICA E TECNOLÓGICA DA UTFPR, 22., 2017, Londrina. **Anais eletrônicos...** Londrina: UTFPR, 2017. Disponível em: <https://eventos.utfpr.edu.br//sicite/sicite2017/index. Acesso em: XXX.
Correspondência:
Lucas Messias Valério
Av Alberto Carazzai, número 455, Bairro Centro, Cornélio Procópio, Paraná, Brasil.