International Journal of Multimedia Information Retrieval (2020) 9:61–62
https://doi.org/10.1007/s13735-020-00194-y

EDITORIAL

Special issue on deep learning in image and video retrieval

Ard Oerlemans[1] · Yanming Guo[2] · Michael S. Lew[3] · Tat-Seng Chua[4]

Published online: 16 May 2020

In recent years, deep learning techniques have been rapidly evolving and advancing, which has resulted in such approaches finding their way into almost all fields where machine learning had already been used or where classic algorithms were still the chosen solution. From image classification and image segmentation to natural language translation and climate modeling, the deep learning approaches have consistently outperformed the best previous algorithms.

This special issue aims to capture the state of the art in deep learning in the context of image and video retrieval. We are pleased to present the following five papers, that were selected after a triple peer-review process:

A useful feature in an image retrieval context is the number of people present in an image, especially when it involves a large number of people. The paper "Single Image Crowd Counting: A Comparative Survey on Deep Learning-based Approaches" by V. Nguyen and T.D. Ngo provides an overview of recent research on crowd counting based on single image input. They start with discussing the approaches that were used before the deep learning era and then show the different aspects of recent research in crowd counting, like multi-scale approaches, divide-and-conquer techniques and coarse counting.

An overview of the state of the art of deep learning applied to video understanding is given by the paper "A Study on Deep Learning Spatiotemporal Models and Feature Extraction Techniques for Video Understanding" by M. Suresha, S. Kuppa and D.S. Raghukumar. With over 100 references, the authors start with a discussion on the extraction of spatiotemporal features from video data and how deep learning models can be used for that. Thereafter, they explore real-world video understanding problems and investigate the future perspective research avenues. This overview provides a general understanding of diverse deep learning strategies for video understanding problems.

In the context of video retrieval, being able to detect sentiments and emotions in video can be used as a powerful feature. The paper "Multi-level Context Extraction and Attention-based Contextual Inter-modal Fusion for Multimodal Sentiment Analysis and Emotion Classification" by M.G. Huddar, S.S. Sannakki and V.S. Rajpurohit presents a multimodal deep learning model for sentiment and emotion classification. The model first extracts the textual, audio and visual features from each of the utterances and then iteratively employs biLSTM with an attention module to extract important contextual information, thereby generating the final trimodal features, which are finally proceeded for classification. By incorporating the contextual information among utterances in the same video, the model outperformed the existing top methods.

The paper "Learning Visual Features for Relational CBIR" by N. Messina, G. Amato, F. Carrara, F. Falchi and C. Gennaro defines Relational Content-Based Image Retrieval as the task of finding images containing similar inter-object relationships. They proposed a relationship-aware deep learning model to extract compact visual features. Given the novelty of the task, the authors employed the CLEVR dataset to generate a relational benchmark. Furthermore, the authors proposed a two-stage Relation Network (2S-RN) module to collect non-aggregated visual features, and its extension module, Aggregated Visual Features Relation Network (AVF-RN), to aggregate the visual features directly inside the network. Both models (2S-RN and AVF-RN) were shown to significantly improve the performance on the R-CBIR task.

✉ Michael S. Lew
mlew@liacs.nl

Ard Oerlemans
ardoerlemans@google.com

Yanming Guo
guoyanming@nudt.edu.cn

Tat-Seng Chua
chuats@comp.nus.edu.sg

[1] Google Research, Mountain View, USA

[2] National University of Defense Technology, Changsha, China

[3] Leiden University, Leiden, The Netherlands

[4] NUS, Kent Ridge, Singapore

Adversarial attacks on deep learning models are well-known method to fool a network by providing inputs that are only slightly perturbed, but provide a completely different output from the model. In the paper "A Retrieval Based Approach for Diverse and Image-specific Adversary Selection" by R.S. Ravat and Y. Verma, a retrieval-based approach is proposed to find adversarial image perturbations that are tailored to the specific query image. Evaluation on the ImageNet dataset and four well-known CNNs showed that their system achieves state-of-the-art performance.