UNDERGRADUATE RESEARCH IN NATURAL AND CLINICAL SCIENCE AND TECHNOLOGY (URNCST) JOURNAL
Read more URNCST Journal articles and submit your own today at: https://www.urncst.com

RESEARCH PROTOCOL OPEN ACCESS

Image Classification by Image Subsets for Fine-Grained Image Recognition

Dario G. Morle, BASc Student [1]*

[1] Department of Electrical and Computer Engineering, University of Windsor, Windsor, Ontario, Canada

*Corresponding Author: morled@uwindsor.ca

Abstract

Fine-grained image recognition is a problem in Computer Vision which focuses on discriminating between objects that appear similar. Two images of an object in this problem can appear vastly different while images from different classes can appear nearly identical. To solve this problem, one must determine regions of significance or salient regions of this image and determine the classification from these regions. Current approaches take the approach of a hard extraction of these regions or some small deviation off hard extraction. Furthermore, in most cases, these regions are used without the spatial context of the region positioning in the image, using an approach similar to the bag-of-words model found in natural language processing. The approach described in this paper will abandon salient region proposals, electing instead to decompose the image into a series of subsets. Each of these subsets will undergo the same feature extraction process, carried out by a series of convolution and pooling layers. The output of this process will be used as the input to a recurrent neural network, ultimately classify the initial image. In processing the image in this fashion, each of these subsets' salience in the context of the larger classification will be determined. A standardized implementation of this architecture has not yet been completed. As such, results indicative of performance can not currently be determined.

Keywords: image classification; image subsets; fine-grained image recognition

Introduction

A classical problem in the field of Computer Vision is image classification. That is, for example, given an image of a dog, determining whether a dog is within the image. This is a problem which for humans is trivial but is not for a computer. A solution to this problem would have numerous real-world problems from facial recognition, to industrial automation and even self-driving cars. Fortunately, the convolutional neural network (CNN) is very good at solving many of these classification problems. But some are beyond the scope of the CNN, one of which is fine-grained image recognition. This problem suffers from large inner-class variance and small intra-class variance.

Current approaches to the fine-grained image recognition problem take the approach of extracting salient regions of the image and largely classifying the overall image entirely on these regions [1].

Bilinear Convolutional Neural Networks (B-CNN) function by passing the image through two independent CNNs and taking the tensor product of the outputs as the input to the fully connected layers for classification. This design is effective because one network can run feature extraction while the other determines salience. The tensor product will then only allow the salient feature into classification [2].

Multi-Attention Networks (MA-CNN) take the approach of running the image through a series of convolutions. Then the output is passed through parallel SE blocks which are each responsible for a salient region of the image. Thereby, aggregating the results will produce an accurate classification based solely on salient regions [3][4].

A more explicit extraction of salient regions is found in the array of models which uses a Region Proposal Network (RPN) to determine possible regions of salience from which the classification is determined [5]. A model which implements this approach is the Object Part Attention Model (OPAM) which annotates each of the proposed regions before classification [6]. Although these approaches are effective, the datasets required must not only contain labels for images, but also annotated bounding boxes for salient regions, making this approach unsuitable for many applications.

Methods

The method which this paper proposes consists of first decomposing an image into a series of subsets. Then, each of these subsets would undergo feature extraction via a common Convolutional Neural Network (CNN). Then, each of these outputs would be analyzed sequentially by a Long Short-Term Memory Recurrent Neural Network (LSTM),

which would output the classification confidence values for each label as seen in Figure 1.

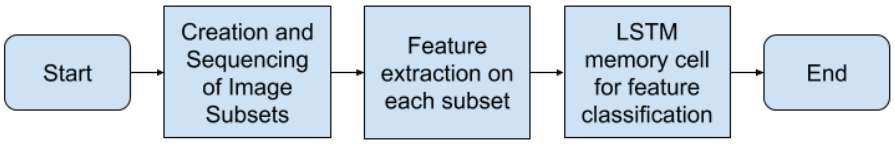

Figure 1: Overview of proposed image classification method

Image Subsets

To start this process, the given image must be decomposed into a series of image subsets to allow for individual analysis of image components. This process can be thought of as finding a mapping from the two-dimensional space of the image's spatial dimensions down to a single dimension containing subsets of the image such that it can eventually be analyzed sequentially. In addition to this, a restriction can be added which will help select a mapping. For the feature classification to work optimally, it would be best if small deviations in the two spatial dimensions of the image correspond to a similar deviation in the sequence of subsets generated. In mathematics, such a mapping exists in the form of space-filling curves. One such curve which satisfies the needs of this process is the Hilbert Curve. The decomposition of one of the Caltech 256 images along a level 1 Hilbert Curve can be seen in Figure 2.

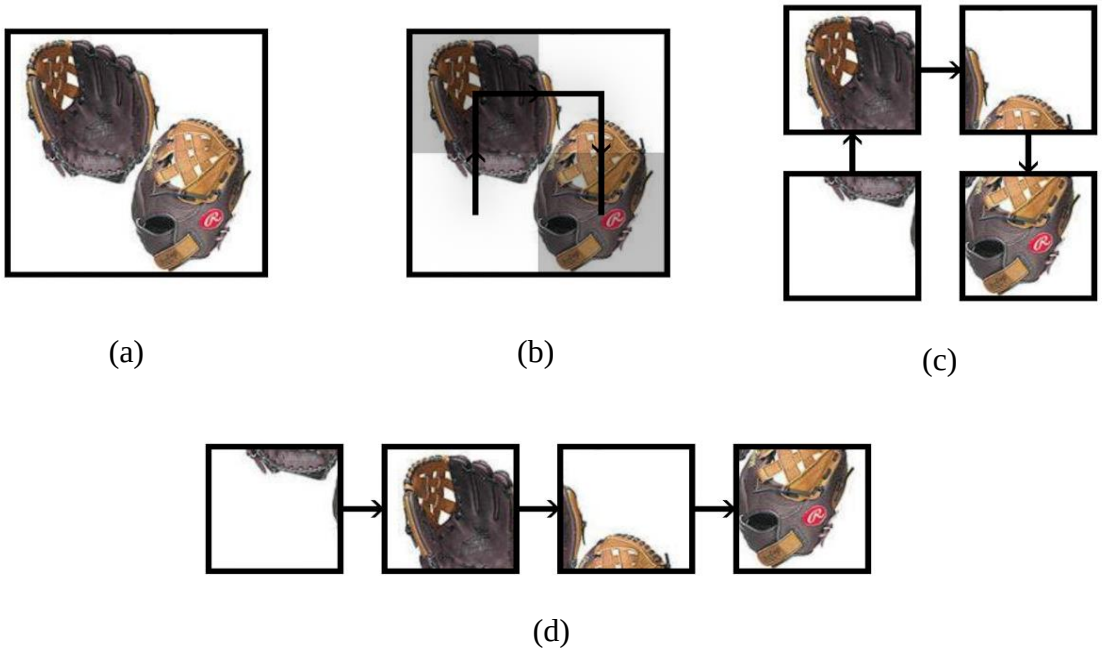

Figure 2: Level 1 Hilbert Curve decomposition of an image in the Caltech 256 Image database [7].

For the purpose of this network, a higher-level Hilbert Curve would be suggested, resulting in more image subsets.

Although Figure 1 displays an image decomposition process during which a pixel was never reused, this is not mandatory. In fact, to minimize the probability that a salient region is not divided along a subset boundary, assuming optimized scaling, it would be best to have any two adjacent subsets have exactly half overlapping. It is also worth noting that an overlap of adjacent subsets greater than half the size of a subset will result in redundant computing.

Feature Extraction

To perform feature extraction, a standard convolutional neural network was used across all image subsets to create a 100-dimensional feature space for each subset as seen in Figure 3.

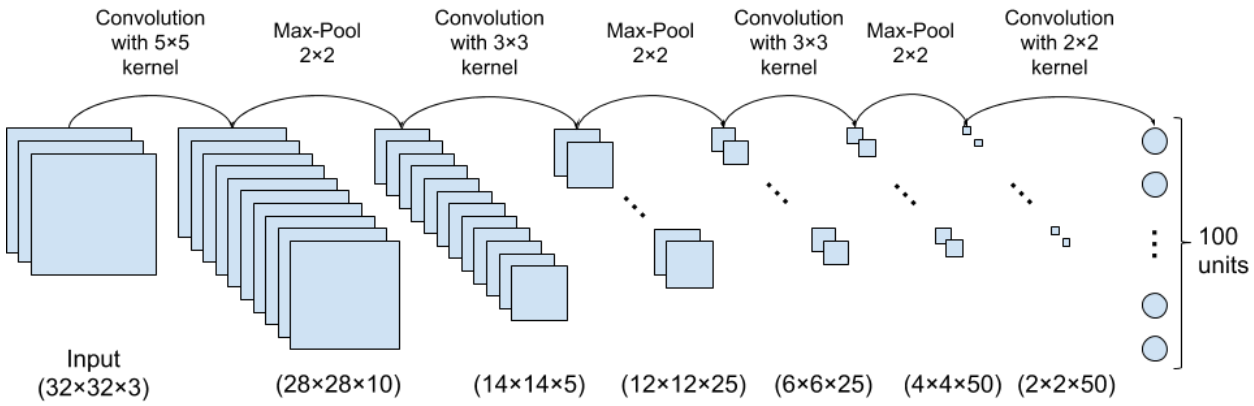

Figure 3: Standard CNN structure with ReLU activation function

In this model, the expected behavior, is that through training, the idea of salience will be encoded into the feature space in such a way such that when the results of this feature extraction process are passed into the feature classification section, the current prediction will be adjusted accordingly.

Feature Classification

Thus far in the image classification process, each image subset created in the first step has been processed entirely independent of one another, allowing for parallel computing as well as region-based features to be extracted. It is in this step where the information between subsets is combined to create a final prediction according to Figure 4.

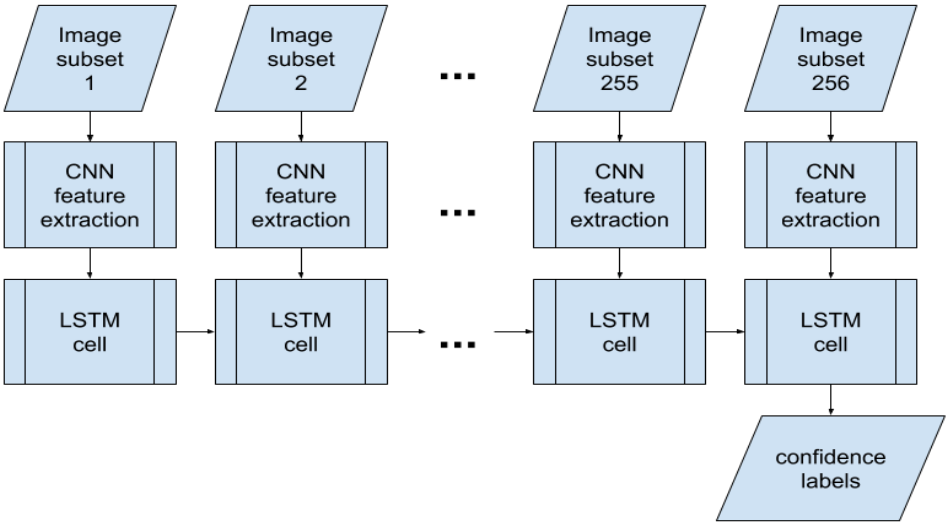

Figure 4: Dataflow regarding feature classification section of the proposed design

In the classification process, a standard LSTM recurrent neural network was chosen specifically because of its ability to propagate information over many iterations if necessary. This will be especially useful given that only a small portion of the image subsets will contain salient information from the feature extraction, and therefore relevant information may need to propagate over many iterations.

Conclusion

This proposal is still a work in progress as a standardized implementation has not yet been completed. Despite this, there are still improvements which can be made to this network which will eventually be implemented. As the proposed network would rely on region salience being extracted alongside features in the convolutional neural network, including a series of SE blocks within the architecture would most likely increase the network's ability to extract salience in the form of insignificant regions being mapped to a channel of zeros [8].

The future of this proposal will take the form of implementation and varying selected parameters of the network under various conditions. For example, with the Caltech bird classification dataset, a finer Hilbert Curve might be necessary for the image subsets to be able to capture

the fine details required in distinguishing bird species whereas in most other datasets a much more course curve could be selected.

List of Abbreviations
B-CNN: Bilinear Convolutional Neural Network
CNN: Convolutional Neural Network
LSTM: Long Short-Term Memory Recurrent Neural Network
MA-CNN: Multi-Attention Convolutional Neural Network
OPAM: Object Part Attention Model
RNN: Recurrent Neural Network
RPN: Region Proposal Network
SE block: Squeeze and Excitation block

Conflicts of Interest
The author declare that they have no conflicts of interest.

Ethics Approval and/or Participant Consent
This study did not require ethics approval or participant consent.

Authors' Contributions
DC: made contributions to the design of the study, collected and analysed data, drafted the manuscript, and gave final approval of the version to be published.

Acknowledgements
The first revision of this paper was supervised by Dr. Thangarajah Akilan of the Computer Vision and Deep Learning Centre at the University of Windsor.

Funding
This study was not funded.

References
[1] Xu H, Qi G, Li J, Wang M, Xu K, Gao H. IJCAI. 2018;:1043-9. Available from: https://doi.org/10.24963/ijcai.2018/145
[2] Lin T, RoyChowdhury A, Maji S. ICCV. 2015;:1449-57. Available from: http://doi.org/10.1109/ICCV.2015.170
[3] Zheng H, Fu J, Mei T, Luo J. ICCV. 2017;:5219-27. Available from: http://doi.org/10.1109/ICCV.2017.557
[4] Sun M, Yuan Y, Zhou F, Ding E. ECCV. 2018;11220:834-50. Available from: https://doi.org/10.1007/978-3-030-01270-0_49
[5] Zhang N, Donahue J, Girshick R, Darrell T. ECCV. 2014;8689:834-49. Available from: https://doi.org/10.1007/978-3-319-10590-1_54
[6] Peng Y, He X, Zhao J. IEEE Transactions on Image Processing. 2017;27(3):1487-500. Available from: http://doi.org/10.1109/ICCV.2015.170
[7] Griffin, G. Holub, AD. Perona, P. The Caltech 256. Caltech Technical Report.
[8] Hu J, Shen L, Albanie S, Sun G, Wu E. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019:1-1. Available from: http://doi.org/10.1109/TPAMI.2019.2913372

Article Information
Managing Editor: Jeremy Y. Ng
Peer Reviewers: Ikjot Saini, Kalyani Selvarajah, Mahreen Nasir Butt, Pooya Moradian Zadeh
Article Dates: Received Aug 14 19; Published Sep 20 19

Citation
Please cite this article as follows:
Morle DG. Image classification by image subsets for fine-grained image recognition. URNCST Journal. 2019 Sep 20: 3(8). https://urncst.com/index.php/urncst/article/view/154
DOI Link: https://doi.org/10.26685/urncst.154