Research Plan/Project Summary

Dense face detection and improving temporal convolutional networks for automatic image captioning

Nikita Zozoulenko

April 15, 2018

1 Rationale

My work is going to be split into the following three parts: (i) A derivation of the general case for a convolutional neural network (CNN) with an arbitrary input of a tensor of order 4 to be used for my own neural network library implementation, (ii) the systematic construction and evaluation of a fully convolutional model for detecting a variable number of faces for real time video, and (iii) creating and improving temporal convolutional networks (TCN) to reach a new state of the art, and to then apply the model on a more complex task which TCNs have never been used for before, namely automatic image captioning, to try to outperform traditional LSTMs and GRUs.

1.1 Derivation of a CNN

When i first stated out in the field of deep learning, I found that the field lacked fully derived examples of convolutional neural networks. Since the model uses tensors of order 4 to represent activations, they cannot be visualized in an effective manner. This causes educational sources to only explain the simple two-dimensional case, instead of the four-dimensional case which is required for the network to perform well. For instance, once concept which revolutionized the training of CNNs is Batch Normalization, and requires all dimensions (a tensor of order 4) to function. It is therefore of the utmost importance that education resources show derivations of the order 4 case, something which no educational resource does. Only teaching students the simple case of order 2 (batch size of one, one channel, and two spatial dimensions), hinders them from reaching their full potential regarding applying CNNs to a variety of problems.

The aim is to present a clear derivation of the general case for convolutional neural networks with an arbitrary input of a tensor of order 4, including varying strides and zero-padding. The aim is to then use my derivations to implement feed-forward and convolutional neural networks with only elementary math operations in Python and C++ for my own library, to gain an deeper understanding and intuition of neural networks. I want to validate my own implementation by training the models on a basic task of classifying handwritten digits, a task where models can be easily trained on a CPU without GPU-acceleration.

1.2 Dense Face Detection

Conventional face tracking and detection methods are limited to a few or just a single face. The aim of this paper is additionally to construct a dense face detector using a fully convolutional neural network which will be capable of detecting a variable number of faces for real time video. This method should be able to be applied to CCTV or other security systems, or areas such as person tagging on social media and face detection for digital cameras.

There are multiple moderns techniques and network architectures used for multi-class object detection. The aim of this paper is to systematically and empirically investigate how the most successful and popular methods, such as Feature Pyramid Networks, Focal Loss, Online Hard Example Mining, and data augmentation affects the performance of a single class object detector. Since object detectors often are used for specifically multi-class classification and detection, the field lacks structured data and empirical evaluations on how these methods affect binary classification problems (e.g. face or no face).

1.3 Improving TCNs for Image Captioning

Additionally, the purpose of this paper is to improve temporal convolutional networks (TCN) in two areas which I have found lacking in the original authors implementation. Firstly, their Pytorch implementation of TCNs is slightly computationally inefficient due to their zero-padding scheme. I'm going to propose my zero-padding method, which final result is mathematically equivalent to their implementation, but is slightly more computationally efficient, while still making sure there's no information leakage from the future into the past. Secondly, I will propose a new dilation scheme which takes inspiration from how conventional spatial state of the art CNNs function. These changes should increase both the efficiency and accuracy of the models, and I will empirically evaluate them on two task. Furthermore, I will compare three different residual building blocks to see their effect on the performance and accuracy of the model.

Improving this sequence model, which has recently been found to perform better than its recurrent counterparts, has an effect on every sequential task. Showing empirical evidence of how TCNs perform better than GRUs and LSTMs, which are traditionally used for sequential tasks, lies in everyone's best interest to further the state of the art in sequential modeling. This can lead to the eventuality of replacing every recurrent network with its temporal convolutional counterpart, while increasing the accuracy on any task.

This is why I'm going to further the research into TCNs by applying them to a more complex task which TCNs have never been applied to before, namely automatic image captioning. By showing that TCNs also outperform GRUs and LSTMs on more complex tasks, a full transition from traditional recurrent network to temporal convolutional networks can be made.

2 Research Questions

3 RESEARCH QUESTION(S), HYPOTHESIS(ES), ENGINEERING GOAL(S), EXPECTED OUTCOMES

4 Procedures

5 Bibliography