UNIVERSITY OF IDAHO, MOSCOW, IDAHO

APPLICATIONS OF CONVOLUTIONAL NEURAL NETWORKS

By

ROSEBELLA CAPIO

A REPORT ON THE APPLICATIONS OF CONVOLUTIONAL NEURAL NETWORKS

November 15, 2018

0.1 Introduction

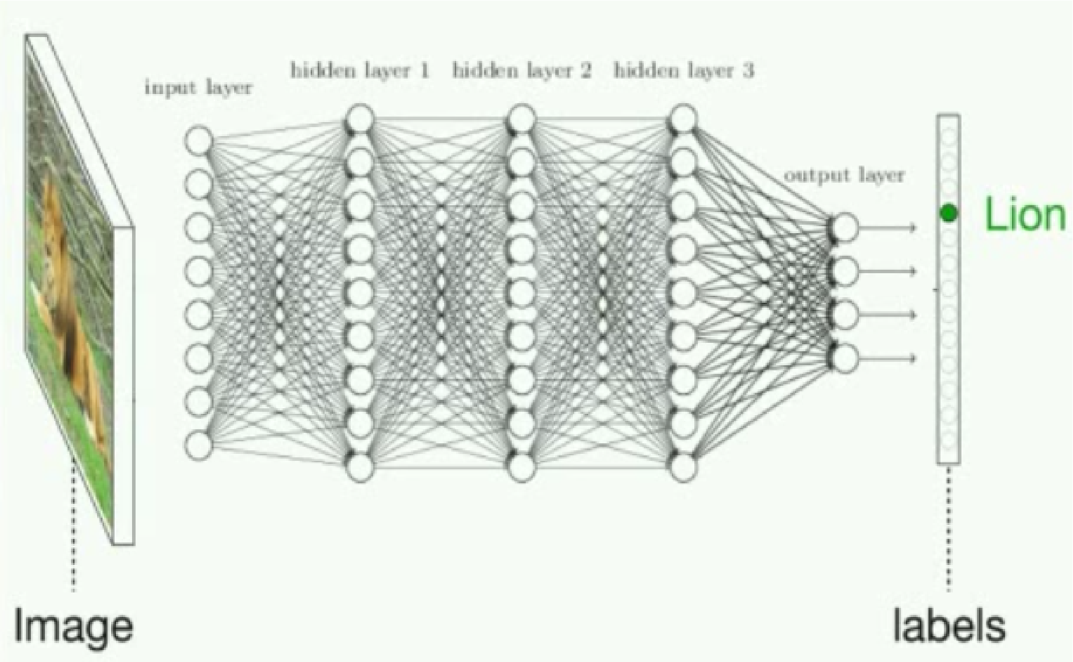
Deep Learning (DL) is a machine learning technique that teaches computers to do tasks that would normally require human intelligence. The architectures or methods use neural networks which are often referred to as Deep Neural Networks due to the number of hidden layers in the neural network. It uses artificial networks and algorithms, which are inspired by the human brain to learn from large amounts of data. DL is the key technology behind driverless cars, medical research (example, cancer cell detection), industrial automation (e.g. improving safety around heavy machinery), and electronics (that is, home devices that can respond to voice, e.g. Alexa)

There are different types of DL architectures. The major differences between the various architectures are basically the structure/ set up of their respective artificial neurons. Some of the types are;

- Recurrent Neural Networks (RNN)
- Recurrent Neural Networks (RNN)
- Convolutional Neural Networks (CNN)
- Deep Stacking Networks (DSN)
- Deep Belief Networks (DBN)

0.2 Convolutional Neural Networks (CNN)

Convolutional Neural Networks is a multilayer neural network that was biolomagically inspired by the animal visual cortex. They are trainable multistage architectures with each stage consisting of multiple layers. The CNN architecture differs from the traditional multilayer perceptrons to ensure some degree of shift and distortion invariance. CNNs convolve learned features with input data, uses 2D convolutional layers, making it well suited to process 2D data, such as images. The image it then divided into receptive fields that feed into a convolutional layer, which then extracts features from the input image. An example of a CNN for an image is seen below;



The input and output of each stage as seen above, are sets of arrays called feature maps. The output stage represents features extracted from all locations of the input. Each stage of the CNN architecture consists of a convolutional layer, non-linearity layer and a pooling layer. The CNNs work by extracting features directly from images. They learn to detect different features of an image by using tens/hundreds of hidden layer with each layer performing a specific function or task (example, the first hidden layer could learn how to detect edges, etc.)

0.3 Applications of CNN

The applications of CNNs looked at in this article fell into two broad categories:

- Computer Vision
- Natural Language Processing.

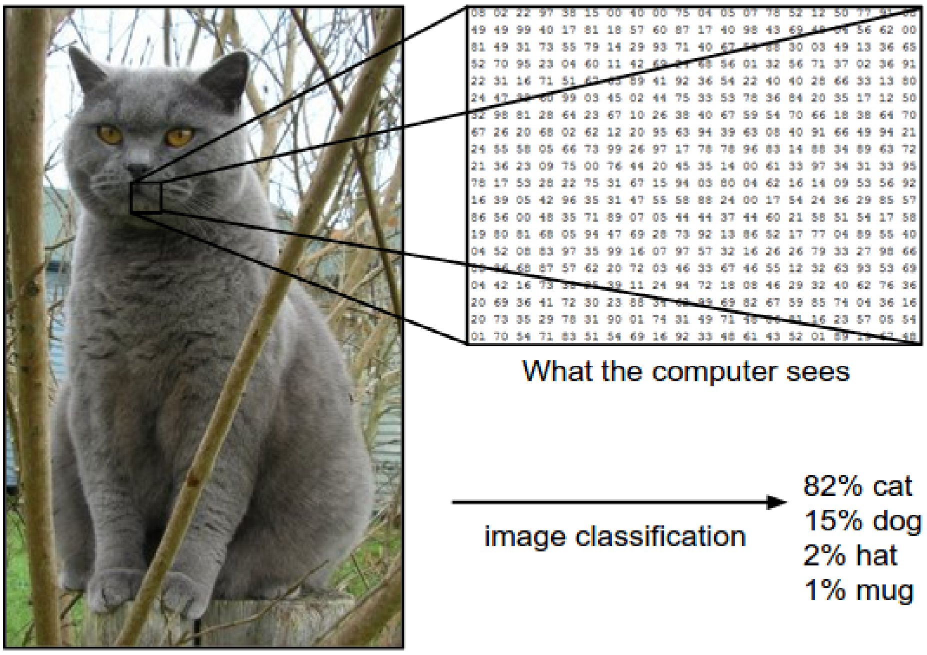
0.3.1 Computer Vision

Convolutional Neural Networks are employed to identify the hierarchy or conceptual structure of an image. A full and complete image is not fed directly into the neural network as one grid, instead, it is broken down into overlapping image tiles where each are fed into a small neural network.
In computer vision, CNN is used for the following;

- Face Recognition
- Scene Labelling
- Image Classification
- Action Recognition
- Human Pose Estimation
- Document Analysis

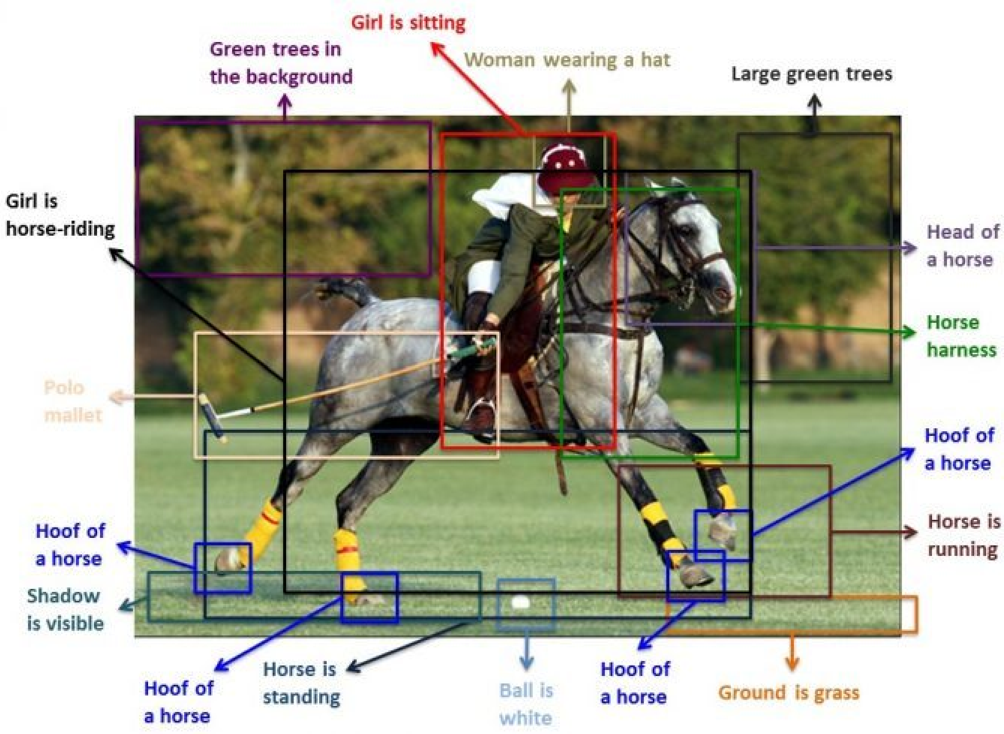
Face Recognition: Faces represent a complex, multi-dimensional visual stimulus. A hybrid neural network combining local image sampling, a self-organizing map neural network and CNN were used to present the complex nature of faces. From the paper, the results were presented using Karhunen-Loe've transform in place of the self-organizing map which performed almost as well (5.3% error versus 3.8%) and a multi-layer perceptron which performed poorly (40% error versus 3.8%). Problems identified in the paper that constitute face recognition are; Identifying all faces in a picture, Focussing on each face despite bad lighting of different pose, Identifying unique faces and Comparing identified features to existing database and determining the person's name.

Image Classification: Image Classification is the task of assigning an input image from a set of categories (example of set of categories can be (dog, cat, mug, hat)). The image classification for this set is given below.



Compared with other models, CNNs achieve better classification accuracy on large scale datasets due to their capability of joint feature and classifier learning. Following the success of AlexNet developed by Krizhevsky et al., which performed really well, several works have been made in an attempt to improving classification accuracy by reducing filter size or expanding the network depth. A GPU implementation of CNN published benchmark resulted in object classification with error rates 2.53% and 19.51% for NORB and CIFAR10. Hierarchical Deep Convolutional Neural Networks (HD-CNN) with CIFAR100-NIN building block was seen to show a testing accuracy of 65.33% which is higher than the accuracy for deep models.

Scene Labelling: Scene Labelling is a challenging classification problem where each input image requires a pixel-level prediction map. Each pixel is labelled with the category of the object it belongs to in scene labelling. The figure below, illustrates schematically a girl on a horse, representing an example of scene labelling.



Action Recognition: Action recognition task involves the identification of different actions from video clips (a sequence of 2D frames) where the action may or may not be performed throughout the entire duration of the video. The difficulties in developing an action recognition system are to solve the translations and distortions of features in different patterns which belong to the same action class. Independent Component Analysis, algorithm to learn invariant spatio-temporal features from unlabelled video data applied on the Hollywood2 and YouTube action datasets gave classification accuracy of 53.3% and 75.8% respectively, which was approximately 5% better than previously published results.

Human Pose Estimation: Huaman-pose recognition is a long-standing problem in computer vision primarily because of high dimensionality of the input data and the high variability of possible body poses. Traditional approaches to human pose recognition depend on appearance cues to predict the human pose rather than motion-based features.

0.3.2 Natural Language Processing

Recurrent Neural Networks (RNN) are generally applied to solve NLP problems due to its resemblance to how human beings process language. But more recently, CNNs have been used in solving such NLP problems like sentiment analysis and spam detection. Some NLP tasks addressed in the paper are;

- Speech Recognition
- Text Classification

Text Classification: Text Classification is an example of supervised machine learning task since a labelled dataset containing text documents and their labels is used for train a classifier. The goal of Text Classification is to automatically classify text documents into one or more defined categories. Some examples of text classification are; detecting spam and non-spam emails, categorization of news articles into defined topics. NLP tasks deals with sentences and documents which are represented as a matrix at the input. Each row of a matrix corresponds to a token, which essentially is a word of in some approaches a character. In the work of Yoon Kim, a CNN-non-static model gave competitive results and improved perfomance by almost 2% with respect to other models which are mainly based on RNN, Autoencoders and Support Vector Machines (SVM).

Speech Recognition: CNN have been used in Speech Recognition recently and have been found to give better results over Deep Neural Networks (DNN) particularly in the areas of Noise Robustness, Distant Speech Recognition, Low-footprint models and Channel-mismatched training-test conditions. In a research done by the Microsoft Corporation in 2015, it was found that CNN obtained relative 4% Word Error Reduction (WER) when trained on 1000 hours of Kinect distance, over DNN trained on the same size. Robustness is also enhanced when pooling is done at a local frequency region and over-fitting is avoided by using fewer parameters to extract low-level features. Pawel Swietojanski et al.

2014 found that WER is improved using CNN by 6.5% relative to DNN for distance speech recognition and 15.7% over a Gaussian Mixture Model.

0.4 Conclusion

CNN is found to give more accurate results in comparison with other traditional methods such as Recurrent Neural Network and boosts performance due to unique features such as shared weights and local connectivity.