PH.D. RESEARCH PROPOSAL

SEQUENCE LEARNING USING DEEP NEURAL NETWORKS WITH FLEXIBILITY & INTERPRETABILITY

Chang Li

Supervisor
Prof. Dacheng TAO

1 Aims & Objectives

1. Flexibility in modeling complex patterns with long-range dependencies

    (a) Capturing complex non-linear correlations without prior knowledge and assumptions
    (b) Encoding high dimensional input variables adaptively
    (c) Discovering long term dependencies of encoded inputs
    (d) Novel insights in feature extraction of complex inputs

2. Network architectures with better computational properties

3. Encoding/Decoding human intepretable representation into/from models

    (a) Learning deep neural networks coupled with structured (potentially predefined) latent variable sequence
    (b) Approximating deep neural networks' input-output relationships using human intepretable models

2 Synopsis

3 Introduction

One interesting task in machine learning is modeling complex sequences having long-term dependencies. Many applications such as machine translation, complex dynamical system analysis, activity recognition and behavioral phenotyping tools for neuro-science involved with capturing non-linear patterns in sequences. Sequence learning mainly have three difficulties: approximating non-linear relationship among sequences, feature selection and capturing long-term dependencies. Our primary aim of this proposal is focusing on demonstrating those three difficulties.

Despite substantial effort has been made for modeling sequences, many of those models are neither unable to approximate non-linear relationships nor have rigid assumptions due to the dependency on predefined form of prior function. For example, autoregressive moving average (ARMA) model [2] and many of its variants [1] have raised interests because of their effectiveness in many real world applications. However, those models cannot capturing non-linear relationships. Probabilistic Graphical Models [3**?**] (PGM) are very well studied for the past decades

4 Expected Research Contribution

5 Proposed Methodology

5.1 Source Modelling

The first step would be source extraction using the standard technique of gaussian modelling. We propose to do this using the robust PyBDSM pipeline used for fitting gaussian distributions to radio sources. The software contains a plethora of features, from which we would be using a small subset. This would mainly include:

1. Source extraction using gaussian modelling of radio data.

2. Generation of a catalog file containing details of radio sources (RA, DEC, Size of Gaussian (min, max), etc.)

5.2 Cutout Generation

The second step would be to convert the RA(Right Ascension) and DEC (Declination) values generated from the catalog, to their corresponding pixel values in the original image. Based on these pixel values we generate 10*10 px cutouts using as reference the co-ordinates of the center of the radio source. This involves a multistep procedure briefly including:

1. Reading the FITS image in the form of a matrix

2. Parsing through the generated catalog file and extracting data for each radio source such as RA, DEC, etc.

3. Converting the RA and DEC values from WCS (World Coordinate System) to pixel values.

4. Processing pixel values to account for difference in addressing between FORTRAN and C family of languages.

5. Slicing the image matrix assuming the reference pixel co-ordinates as the center of the source.

Prototype code for section *3.1* and section *3.2* has been written mainly for testing purposes. We used a sample image from the TGSS survey which was then processed using the first two steps of our pipeline to generate 470 cutout images. More details can be found at: https://github.com/NCRA-TIFR/radiogen.

5.3 Data Preprocessing

Real world data is incomplete, inconsistent and noisy. Techniques like data cleaning, integration, transformation, reduction, discretization are required to ensure data uniformity.

Standard techniques include:

1. Mean Subtraction

   The mean is subtracted across every feature in the data cloud. The geometric interpretation of this operation is the centering of the data cloud around the origin along every dimension.

2. Normalization

   The data dimensions to span the same range. This results in higher accuracy during data processing, as inter-dimensional variance is accounted for. There are two common ways of achieving this normalization.

   (a) Divide each dimension by its standard deviation, once it has been zero-centered.

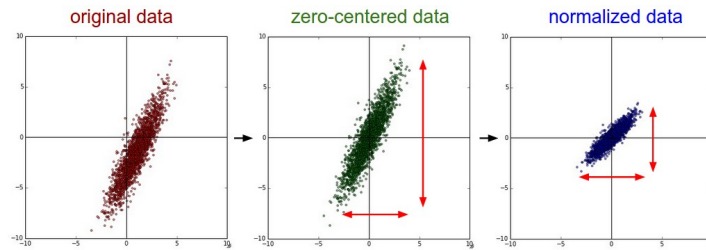   (b) Normalize each dimension to result in a range of values [-1,1].

Figure 1: Mean Subtraction and Normalization

3. Dimensionality Reduction using Principal Component Analysis

   The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. By finding the eigen vector with the highest eigen value we select the Principal component axis with the highest variance and the minimum reconstruction error. Dimension reduction occurs by eliminating axes with eigen vectors exhibiting the least variance. This results in a data representation exhibiting a higher degree of variance, leading to faster convergence and greater accuracy during classification.
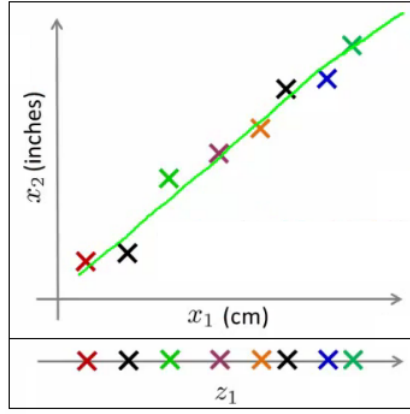
Figure 2: Reduction from 2D to 1D using PCA

5.4 Analytical Approach

The analytical approach to solving the problem of galaxy morphology classification would primarily contain two steps:

1. Feature extraction using sophisticated techniques prevalent in image processing.

2. Building a statistical model of the extracted features

3. Cross comparison of the designed model with principal indicators of dominant classes

A variety of techniques have been used to tackle the related problem of visual recognition in Computer Vision. The standard algorithm for feature extraction is the Scale Invariant Feature Transform (SIFT).

5.4.1 Scale Invariant Feature Transform

The SIFT algorithm extracts keypoints from a set of reference images and stores them in a database. The feature vectors of the extracted images are compared wih the reference vectors on the basis of a defined norm (usually Euclidean distance). The probability that of the subset of features indicating the presence of an object is computed. This gives the accuracy of the fit. When multiple features pass these tests, the class can be correctly identified with a high confidence.
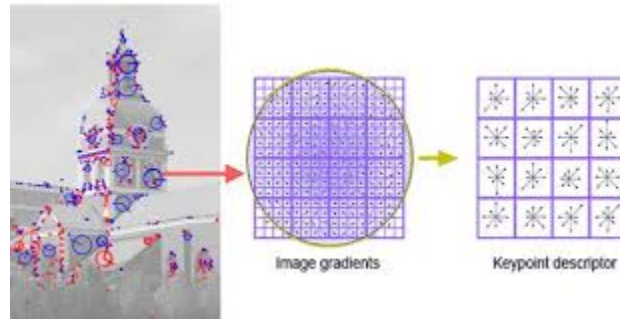

Figure 3: Feature Extraction using SIFT

5.5 Empirical Approach

An empirical approach is a way of gaining knowledge by means of direct and indirect observation or experience. In the context of this project, we plan to employ Artificial Neural Networks(ANNs), which are a computational model based on a large collection of artifical neurons, that learn hierarchical representations of the data. Each neural unit is connected with many other units which computes to essentially create a non-linear stacked function approximator. Neural networks typically consist of multiple layers in which a signal traverses from the input layer to the output layer.

A convolutional neural network (CNN), is a type of ANN, is a state of the art model that currently provide the best solutions to many problems in computer vision such as image segmentation, recognition and classification[?].

Convolutional neural networks were designed to use minimal amounts of preprocessing. A typical architecture of a CNN consists of convolutional layers, activation layers and pooling layers each of which is briefly explained.

- Convolution layer: Performs the convolution operation with a static filter on input from previous layers.

- Activation layer: Defines the output of a unit given an input or set of inputs by applying an activation function.

- Pooling layer: Progressively reduces the spatial size of the representation, which results in a decreased number of parameters and thus reduction in network computation.

The initial layers of the CNNs learn simple features in the input data such as straight edges, simple colors, and curves. The deeper layers learn the higher level representation of image and search for complex shapes and structures. Mathematically, deeper layers can be thought of as compositions of previous layers. [?][?][?].
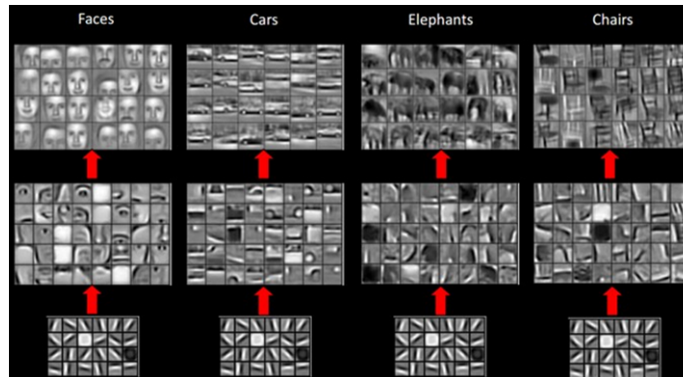

Figure 4: CNN learning simple features on initial layers for different classes

Variants of CNNs have achieved great successes for morphological galaxy classification and prediction on SDSS[?]. Sander Dieleman et al. won the Galaxy challenge, an international competition on Kaggle, to build the best model for morphology classication based on annotated images from the Galaxy Zoo project.

6 Work Plan

April - May ─● Literature survey
Mid-August to September ─● Hypothesis testing
September to October ─● Prototype implementation
October to November ─● System refinement
November to December ─● Publishing results

7 Conclusion

References

[1] P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.
[2] M. Hibon and S. Makridakis. Arma models and the box–jenkins methodology. 1997.
[3] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

---

[1] National Center for Radio Astrophysics - Tata Institute of Fundamental Research