CS 771A: Intro To Machine Learning

Handwritten Mathematical Expressions to LaTeX

Group: 27 Abhyãsa

Aravind reddy{arareddy@iitk.ac.in} Gunda Abhishek{abhigun@iitk.ac.in}
Harsha Nalluru{harshan@iitk.ac.in} Krishna Karthik{jkrishna@iitk.ac.in}
Gowtham Prudhvi{gowthamp@iitk.ac.in}

September 21, 2017

Problem Statement

Converting handwritten mathematical expressions to LaTeX code using Machine Learning.

Mathematical expressions are a critical part in most engineering disciplines. Giving a mathematical expression as an input to a computer is more troublesome when compared to giving plain text. LaTeX is very flexible and powerful tool which has the ability to render complex mathematical expressions. In this project, we aim to build an application which is capable of taking a handwritten mathematical expression as input and outputs the LaTeX code for the corresponding expression.

Approach

An application is to be built where input from the user is taken from a sketch pad, application takes snapshot of this handwritten expression and gives as an input to the machine learning model. The following steps are performed to the input.

Expression Detection

Expressions must be first identified and segmented. Segmentation[1] gets more difficult as there would be distinctions in the size of symbol which is quite a challenge to extract the symbols from an expression.

Symbol Recognition

Objects obtained from the segmented expression are tested for the character they represent. Variations in the writing styles and the image quality makes it more difficult to recognize the symbol. It requires the machine learning model to surpass these difficulties.For each of the characters, the model will be generating scores for the corresponding results.

Latexification

Based on the scores given by the machine learning model for each of the characters in the expression, the resulting expressions will be ranked in the order of resemblance with the original expression. The latex code of the expression that is highly likely to resemble the input is provided as the output of the application.

Machine Learning Model

Designing an algorithm to extract characters from the expression and developing an ML model for recognizing the symbols using the Handwritten math symbols dataset[5] from Kaggle[5], by training and testing using the Nearest Neighbour algorithm[2], SVM[3], Neural Networks, e.t.c., algorithms, and eventually integrating it in the application's workflow to output the results in the order of resemblance.

Advancements

Extending the application to Online Learning[4], where each of the characters of the input are added to train the model in an online fashion.

Existing work

The links given below are part of the existing work that has been done over the past years .

- The segmentation of images as discussed above can be done using various methods which can be found in [3]. The model was trained using an SVM classifier for which the accuracy for 9 classes chosen on the test data set was 90%. We hope to increase the size of the dataset while maintaining the accuracy.
- [7] addresses the problem of segmenting an image into regions.
- In [1], segmentation is done using the K-Means Algorithm and character recognition is done using Convolutional neural networks.
- Machine Learning model trained using neural networks, the accuracy for 101 classes chosen on the test data set is 81.5%. We hope to increase the size of the dataset and improve the accuracy. [8]
- OpenCV's OCR of Hand-written Data of characters in a single image using kNN.[2] This might partially help in character recognition after segmenting the image.

References

[1] J. Chang, S. Gupta and A. Zhang. *Painfree LaTeX with Optical Character Recognition and Machine Learning.* Project Report, CS229, Autumn 2016, Stanford University.

[2] A. Mordvintsev and K. Abid. *OCR of Hand-written Data using kNN.* OpenCV- Python Tutorials, Revision 43532856, 2013.

[3] Abirami M, Rajashri S and Ramya R. *Comprehending Handwritten Mathematical Formulation Using SVM Classifier.* International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue. 3:4556-4560, March 2017.

[4] M. Thoma. *On-line Recognition of Handwritten Mathematical Symbols.* arXiv:1511.09030, 2015.

[5] X. Nano *Handwritten Math symbols dataset.* Kaggle

[6] T. Lech *Math to Latex Blog*

[7] Felzenszwalb, Pedro F. and Daniel P. Huttenlocher. *Efficient Graph-Based Image Segmentation.* International journal of computer vision 59.2 (2004): 167-181.

[8] Y. Peng and Y. Zhang. *Offline Mathematical Character Recognition.* Project Report, CSCI B657, Spring 2016, Indiana University.