Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-2 Issue-5, August 2022

Time-Efficient Algorithm for Data Annotation using Deep Learning

Sumit Chaudhary, Neha Singh, Salaiya Pankaj

Abstract: ***Current generation emphasis on the Digital world which creates a lot of unbeneficial data. The paper is about data annotation using deep learning as there is a lot of data available online but which data is useful can be labeled using these techniques. The unstructured data is labeled by many techniques but implementation of deep learning for labeling the unstructured data results in saving the time with high efficiency. In this paper introduce a method for data annotation, for that we can use unlabeled data as input and it is classified using the K-Nearest Neighbor algorithm. K-Nearest Neighbor is the fastest and its accuracy is very high compared to other classification algorithms. After classified the unlabeled data we use it as input and annotate data using deep learning techniques. In deep learning we use an auto annotator for annotating data. After annotating data, check the accuracy of annotated data and time efficiency of data annotation. In case the accuracy is low then we can retrain the data and make it more accurate.***

Keywords: K-Nearest, Neighbor Algorithm, Deep Learning

I. INTRODUCTION

A Neural network consists of three or more layers in which machine learning consist of deep learning and then Artificial Intelligence consists of both machine learning and deep learning. To simulate the activity of the brain of humans we require neural network to match the skills to learn from big amount of data. [1]

Artificial intelligence applications for data are classification and tagging using data annotation. The specific condition training data must be classified and annotated in such a good way. For AI operation and build a high performance for such a task with best quality and human powered data annotation. The Data annotations are classified in many types like text annotation, audio annotation, image annotation, and video annotation. [2]

Manuscript received on 30 July 2022 | Revised Manuscript received on 13 August 2022 | Manuscript Accepted on 15 August 2022 | Manuscript published on 30 August 2022.
***** Correspondence Author
Dr. Sumit Chaudhary, Associate Professor & HOD, Department of Computer Science and Engineering, Indrashil University (Cadila Group), Kadi (Gujarat), India.
Ms. Neha Singh*, Assistant Professor, Department of Computer Science and Engineering, Indrashil University (Cadila Group), Kadi (Gujarat), India.
Salaiya Pankaj, Department of Computer Science and Engineering, Indrashil University (Cadila Group), Kadi (Gujarat), India.

A. Database Creation

Database creation it thirty-six participants take part in study. From that every participant completes ten one-hour driving periods. Participants were assigned a task to drive to a certain area after obtaining practice drives to grow acquainted with the simulated driving environment. The simulated environment was created for the participants which is as familiar and natural as possible, and this simulated world was designed to closely resemble the local metropolitan area such that navigation assistance is not required by the participants for accomplishing their aim. Traveling to workplace, then traveling to pick a friend from airport from workplace, traveling for lunch and in the end back to home from workplace were among the tasks. The only other direction given to participants was to drive normally. Each drive was around 10 and 25 minutes long. Participants drive several times during each session as time permitted. This design includes two key elements that promote increased realism in driving and, as a result, in the data collected: comfort with the driving task, and involving participants in the activities. The test resulted in 132 hours of driving time and about 265 GB of data was collected. [3]

We gather information together in a driving simulator lab that consists of an experimental test car in a virtual environment with full visual and audio simulation models of various roads, traffic, and road activity. A resolved automobile is encircled by five forward and triple rear displays in the driving simulator. All driving inputs, including the gear stick, brake, as well as accelerator, are followed and effect the virtual world's movements in real time. To simulate genuine driving, several hydraulic systems and motor's structure feedback signals to driver controls. A driving simulator application is a commercial package, a collection of simulated devices that, at the behavioral level, imitate a wide range of existing and upcoming on-board sensors. Multiple video cameras, microphones, and laser eye movement sensors are included in the simulator setup to capture all driver activities throughout the drive, which are synced with all sensor's information and simulator monitoring parameters.[3]

B. Database Collection

A systematic way of collecting data or observation is called Data collection. A big amount of data collection is the overall process of it. For it, first you target the purpose of research, select the useful data and collect, after that you can process the way to gather the data, collect, store and process the data with proper technique.

For that thing you must consider data collection.[4] Qualitative data and Quantitative data are two main types of data. The number of data is not measured but it is good for use that data is called Qualitative data. And the other side the data is measured by the numbers it's called Quantitative data.[5] Time efficiency is a computer science term that quantifies the amount of time it takes a set of code or algorithm to process or run-in relation to the amount of input. To put it another way, the temporal efficiency measures how long it takes a programmed to process a given input. It is the number of times a certain code is performed rather than the overall time consumed. This is because the overall time taken is determined by various elements such as the compiler used, the processor's speed, and so on. [6]

II. LITERATURE SURVEY

A. Time Efficient Algorithm

Data annotation is a use of a large amount of database in any area. The process of data annotation is taking too much time and it will also be expensive. In this survey we study to reduce data annotation time using machine learning techniques. For the reduced data annotation time use Random forests algorithm. In this tool Random forests are used for bootstrapped classifiers. The tool is able to generate annotations and verified automatically by itself. A tool can be able to reduce for one minute of time to around thirty-five seconds annotation time. [3]

The growing use of high-frequency (kHz) intracranial monitoring from numerous electrodes during pre-surgical epilepsy assessment generates vast volumes of data that are difficult to store and retain. Simple, typically manual, data analysis approaches are further challenged by annotations and diagnostic annotations in these massive data sets. The challenges of dependable metadata and annotation transmission across programmed, the preservation of meaning for that particular information with increased time period, and the ability to resorting of data for the purpose of analysis it imposes the data structure and algorithms all different demands. Individual issue domain solutions can be configured to allow for easy interpretation and clarity across domains. [7]

In comparison to "learning from examples," active learning posits that the learning algorithm has some influence over which parts of the initial dataset it gets information about. Active learning is demonstrably more potent than learning from examples alone and in particular instances, resulting in improved generalization for a limited training data set. We define the selective sampling approach for active concept learning and illustrate how it may be approximated by a neural network. A learner obtains statistical information from the context and queries an oracle on areas of the domain that it thinks valuable in selective sampling.[8]

B. Deep Learning Strategies with Data Annotation.

Active Learning can help you save time as well as cost on data annotation. It is a combination of approaches that display the data annotation stages as an interaction among a deep learning model and a user, in which the algorithm recommends which cases are worth annotating while the user annotates the samples picked.[9]

The fast advancement of artificial intelligence, as well as the protection of intellectual property of deep learning models, has scientists and engineers concerned. The majority of trigger sets in literates, on the other hand, were built using understandable elements like Gaussian noise and badges on clothing. Original data source the watermarking characteristics can then be obtained via machine learning attacks. Create a fictitious trigger set as a result, fake property claim assaults are possible. The main purpose of this paper is to focus on data annotation and present a black-box watermarking approach based on unstructured autonomous data annotation. annotated data, the sensitivity of the starting value, aperiodic action, and aperiodic actions are all advantages of chaos.

Unstructured sequence's unpredictability These unstable characteristics are used to data annotation in order to combat the assault on bogus ownership claims to begin, this approach uses unstructured automated data annotation, which saves time and eliminates the need for human labeling.[10]

In a given scenario Optical object recognition aims to identify objects of particular target categories with proper accuracy and then seek a class label to each and every object instance. Object recognition strategies for deep learning have been greatly investigated in recent years as a result of deep learning-based picture classification. In this paper recent modifications and improvements are discussed with an overall overview. The three primary parts in which existing object recognition framework and the survey structure is divided are: (i) detect components, (ii) learning methodologies, and (iii) implementations & benchmarks, after analyzing a significant body of recent relevant work in literature.[11]

High-performance deep learning algorithms now in use rely on huge test dataset with productive user annotations, which are tough to obtain in many clinical applications. Annotation-efficient Deep learning (AIDE), an open-source system for dealing with poor training datasets, can be seen here. We execute systematic analysis and practical assessments to show that AIDE outperforms traditional fully supervised models on open datasets with sparse or messy annotations.[12]

Three datasets comprising pictures from three medical centers were used, and AIDE consistently produced classification maps that were equivalent to those created by fully-supervised equivalents or supplied by independent radiologists using 10% training annotations. The 10-fold increase in expert label accuracy has the potential to benefit a wide range of biological applications. [13]

Due to high labeling costs, few poorly labeled training examples are available, making logo identification in unregulated images difficult.[14] In this paper, we provide a model training image generating approach capable of greatly boosting logo identification accuracy when just a few labeled training photos collected in a practical manner are available, while avoiding the high expenses of human labeling. We propose a unique technique for producing New Context trained photos in order to improve model resilience against unexpected backdrop clutters and thereby improve logo review and evaluation. [15]

C. K-NN Algorithm

The k-Nearest Neighbor (kNN) algorithm is a simple yet effective machine learning method. It works well with both classification and regression. It is, however, more usually applied in classification prediction. kNN classifies newly inputted data based on its similarities to previously trained data and arranges the data into meaningful clusters or subsets.[16] A improved KNN approach for classification tasks has been presented in which the classification model is built by combining KNN categorization with a restricted single pass clustering technique. If all of the classes have a constant value of K, the class with the most characteristics will have an advantage. In enhanced KNN, a reasonable number of closest neighbors is utilized to forecast the class of unlabeled data based on the distribution of data in the training set.[17] The use of shared nearest neighbour similarity to determine the similarity between test samples and nearest neighbour samples improves the K-nearest neighbour method.[18]

III. METHODOLOGY OR ARCHITECTURE

In this paper, there is introduction of a new method of data annotation using deep learning techniques. We can use this method for data annotation which is fast and more efficient. We can use row unlabeled data from the data sets and this data set is classified through the K-NN algorithm. The K-NN algorithm is the fastest and more accurate algorithm for classified unlabeled. After classifying the data, we predict if the data is uncertain or not if data is uncertain then data is sent for data annotation. Here data is for the data annotation process, use the deep learning technique for annotating the unlabeled data. After row labeled the data annotation, the data is gone for collecting the sample labeled data set. The final labeled data is created and after the creation of the final data set, efficiency of the data is calculated and if the efficiency is achieved then it stops, if not then it further goes for the same process until the standard efficiency is achieved.

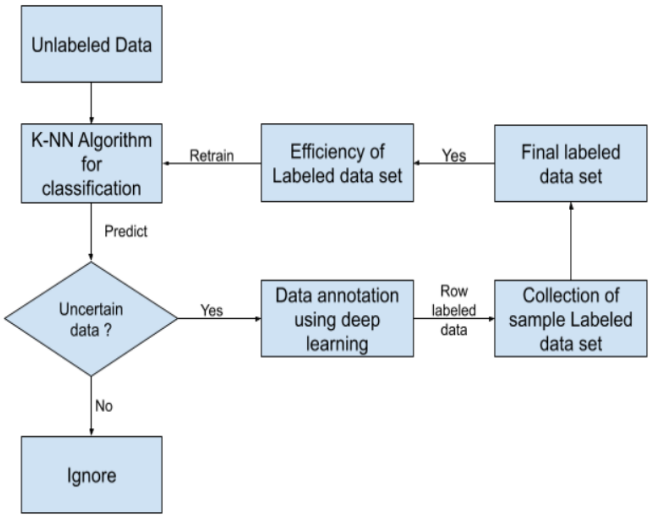
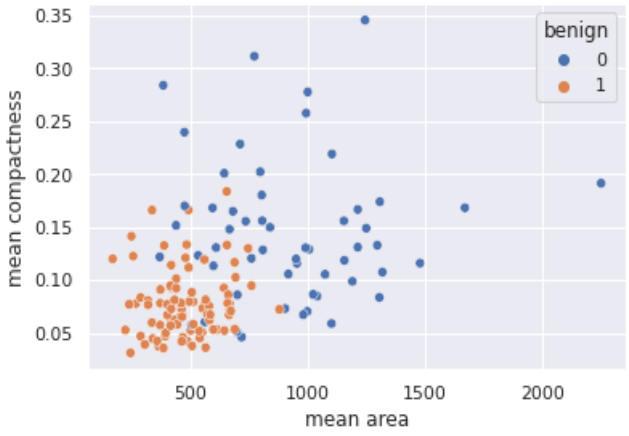
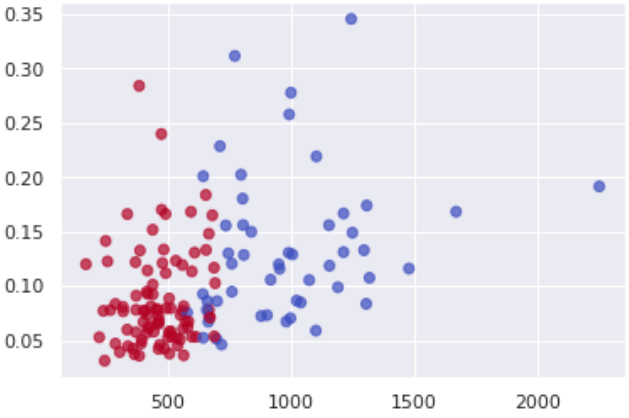


Figure 1: Proposed Architecture

IV. RESULT





V.CONCLUSION

By going through various literature, the result has shown that the deep learning technique for data annotation is more efficient and more accurate than the traditional methods. While using machine learning techniques, the process time for the data annotation is comparatively slow while using deep learning for the same data annotation the process is fast and efficient.

Classifying the unlabeled data through K-nn algorithm and annotating the uncertain data through deep learning. After that we measure the efficiency of annotated data and go for the next process.

FUTURE SCOPE

- In 2026, an expected 464 exabytes of data would be generated everyday throughout the planet, according to Visual Capitalist. Furthermore, according to Global Market Insights, the global market for data annotation tools is predicted to increase at a rate of 40% per year over the next six to seven years, with the automotive, retail, and healthcare industries leading the way.
- Data annotation is a vital and amazing task, especially given the current rate of data development. It will continue to be beneficial in AI and machine learning applications.[19]
- Improving the Quality of Search Engine Results for Multiple User Types Search engines need to provide users with comprehensive information. Their algorithms must process high volumes of labeled datasets to give the right answer to do that. Take, for example, Microsoft's Bing. Since it caters to multiple markets, the vendor needs to make sure that the results the search engine would provide would match the user's culture, line of business, and so on.[20]
- Refining Local Search Evaluation
  While search engines cater to a global audience, vendors also have to make sure that they give users localized results. Data annotators can help with that by labeling information, images, and other content according to geolocation. [20]
- Enhancing Social Media Content Relevance
  Like search engines, social media platforms also need to provide customized content recommendations to users. Data annotation can help developers classify and categorize content for relevance. An example would be categorizing which content a user is likely to consume or appreciate based on his/her viewing habits and which he/she would find relevant based on where he/she lives or works.[20]

REFERENCES

1. "https://www.ibm.com/cloud/learn/deep-learning".
2. "https://appen.com/blog/data-annotation/#:~:text=Data%20annotation%20is%20the%20categorization,build%20and%20improve%20AI%20implementations."
3. C. Schreiner, K. Torkkola, M. Gardner, and K. Zhang, "USING MACHINE LEARNING TECHNIQUES TO REDUCE DATA ANNOTATION TIME."
4. "https://www.scribbr.com/methodology/data-collection/"
5. "https://www.iteratorshq.com/blog/data-collection-best-methods-practical-examples/"
6. https://www.geeksforgeeks.org/time-complexities-of-different-data-structures/ cited on 05/04/2022
7. Mark R. Bower, Ph.D., Matt Stead, M.D., Ph.D.,Benjamin H. Brinkmann, Ph.D., Kevin Dufendach, Gregory A. Worrell, M.D., Ph.D "Metadata and Annotations for Multi-scale Electrophysiological Data"
8. DAVID COHN ,LES ATLAS,RICHARD LADNER ,"Improving Generalization with Active Learning"
9. https://towardsdatascience.com/active-learning-for-an-efficient-data-annotation-strategy-4d007c5d7ed1 cited on 08/04/2022
10. YING-QIAN ZHANG,YI-RAN JIA, XINGYUAN WANG , QIONG NIU,NIAN-DONG CHEN "DeepTrigger: A Watermarking Scheme of Deep Learning Models Based on Chaotic Automatic Data Annotation"
11. Xiongwei Wua , Doyen Sahoo b , Steven C.H. Hoi,"Recent advances in deep learning for object detection"
12. Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu, Jie Chen, Huihui Zhou, Ismail Ben Ayed & Hairong Zheng, "Annotation-efficient deep learning for automatic medical image segmentation"
13. https://www.researchgate.net/figure/Illustration-of-logo-detection-challenges-significant-logo-variation-in-object-size_fig1_322645998
14. Shanshan Wang,Cheng Li , Rongpin Wang,, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu , Jie Chen, Huihui Zhou, Ismail Ben Ayed & Hairong Zheng,"Annotation-efficient deep learning for automatic medical image segmentation"
15. Hang Su, Xiatian Zhu,Shaogang Gong"Deep Learning Logo Detection with Data Expansion by Synthesising Context"
16. Kashvi Taunk,Sanjukta De,Srishti Verma,Aleena "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification"
17. Shengyi Jiang,Guansong Pang,Meiling Wu,Limin Kuang, "An improved K-nearest-neighbour algorithm for text categorization",Expert Systems with Applications,Elsevier(2012).
18. Wei Zheng,HaiDong Wang,Lin Ma,RuoYi Wang,"An Improved k-Nearest Neighbour Classification Algorithm Using Shared Nearest Neighbour Similarity" ,Metallurgical & Mining Industry . (2015), Issue 10, pg. 133-137. 5p.
19. "https://www.naukri.com/learning/articles/data-annotation-definition-types-tools-and-its-future/" cited on 23/04/2022
20. "https://www.techslang.com/definition/what-is-data-annotation/" cited on 23/04/2022

AUTHORS PROFILE

Dr. Sumit Chaudhary is HOD-CSE/Associate Professor in Computer Science & Engineering department at Indrashil University (Cadila Group), Kadi, Mehsana, Gujarat. He has done Ph.D. in Computer Science & Engineering (CSE) from Uttaranchal University, Dehradun (Uttarakhand) in 2018. Dr. Sumit Chaudhary is having more than 12 years of training and teaching experience including 3.5 years of Research experience. He has worked internationally with Ningxia University, China in the department of School of Information Engineering as Associate Professor/ Foreign Faculty. He obtained his M.Tech. (Computer Science & Engineering) with Hons. from Shobhit University and B.Tech. (Computer Science & Engineering) with Hons. from Government college, SCRIET, Meerut. His area of research includes Big Data Analytics, Cloud Computing, Wireless Sensor Network (WSN), Network Security, Neural Network, Artificial Intelligence and MANET (Mobile Ad-Hoc network.

Ms. Neha Singh is Assistant Professor in Computer Science & Engineering department at Indrashil University (Cadila Group), Kadi, Mehsana, Gujarat. She is having 11 years of teaching experience with Research experience & International exposure of Ningxia University China. She is Pursuing PhD from Uttarakhand Technical University, Dehradun. She obtained her M.Tech. (Computer Science & Engineering) from Shobhit University and B.Tech. (Computer Science & Engineering) with Hons. from Government college, SCRIET, Meerut. She has Supervised many Master students in their research. She has published many national and international research paper and conferences (including IEEE & Springer) with 1 Patent. Her area of research includes Wireless Sensor Network (WSN), Network Security, Artificial Intelligence and MANET (Mobile Ad-Hoc network).

Salaiya Pankaj completed his BTech in CSE with specialization in Machine Learning from Indrashil University in 2022. Now he is planning to pursue MTech in CSE. His interest area includes Machine Learning & Artificial Intelligence.