Automatic Generation Image Captions based on Deep Learning and Neural Network

1rd Yashwanth Nalamasa
department of computer Science
University of central Missouri
Lees's Summit, Missouri
YXN77150@ucmo.edu

2rd Sreeja Madhagoni
department of computer Science
University of central Missouri
Lees's Summit, Missouri
SXM58610@ucmo.edu

3rd Tharun Bhukya
department of computer Science
University of central Missouri
Lees's Summit, Missouri
TXB75040@ucmo.edu

4th Shravya Mendu
department of computer Science
University of central Missouri
Lees's Summit, Missouri
SXM57840@ucmo.edu

***Abstract*—We introduce a novel method for automatic image captioning that employs a Recurrent Long Short-Term Memory (R-LSTM) model. Our strategy involves taking an image as input and generating a descriptive caption through a sequence of words. The R-LSTM architecture, known for its capability to model long-term dependencies in sequential data, constitutes the neural network type at the core of our approach.**

The initial step in our process involves the pre-processing of image data and feature extraction. Here, we leverage a Convolutional Neural Network (CNN) to extract crucial features from the image. Subsequently, we delve into the R-LSTM architecture, elucidating the incorporation of attention mechanisms that enrich caption quality by considering contextual information.

To assess the effectiveness of our proposed model, we conduct evaluations using a benchmark dataset and compare its performance against other contemporary approaches. Our findings indicate that the R-LSTM model surpasses these alternatives in terms of caption quality, as gauged by standard evaluation metrics.

Moreover, we draw attention to the challenges inherent in constructing a proficient automatic image captioning system. These challenges encompass the intricacy of generating semantically meaningful captions and the necessity for substantial amounts of labeled data. Despite these obstacles, we posit that our proposed methodology signifies a notable advancement in this field, holding the potential to markedly enhance the accessibility and usability of visual content on the internet.

In summary, we advocate for the R-LSTM model as a promising avenue for automatic image captioning. We anticipate that our research will stimulate further exploration in this domain and that the ongoing refinement of accurate and effective automatic image captioning systems will exert a substantial influence on image understanding, content accessibility, and image search and retrieval.

***Index Terms*—LSTM,CNN,NLP, Dataset,Word Sense Disambiguation ,VGG.**

[1]

Identify applicable funding agency here. If none, delete this.

[1]https://github.com/tharunbhukya/NNDL_FinalProject.git

I. Introduction

Various image sources, such as television, the internet, and news outlets, provide a plethora of visuals. While people can interpret these images even without accompanying descriptions, machines face significant challenges in doing so. For machines to comprehend images, descriptive information becomes essential.

The field of natural scene captioning stands as a prominent area of research, facilitating the generation of descriptions for images in the era of artificial intelligence [1]. This research holds considerable significance for various reasons, with major entities like Facebook and Google utilizing it to discern users' locations, activities, and engage in similar functionalities.

Comprehending images involves the identification of objects, actions, and the relationships between them. Machines face challenges in recognizing subtle elements, like understanding that individuals are waiting for a train even when the train is not visibly present on the platform. It is crucial that any generated sentence is not only syntactically correct but also semantically valid [2]. To achieve a comprehensive understanding of a specific natural landscape, features are extracted. This extraction process broadly falls into two categories: (1) Methods based on Deep Learning and (2) Methods based on Conventional Machine Learning.
Extensive labeled datasets like ImageNet, coupled with the power of deep learning, offer a significant advantage, particularly in the effectiveness of deep convolutional neural networks (CNN). The field of computer vision has witnessed substantial progress through image captioning, enabling computers to perform various tasks such as early education, video tracking, sentiment analysis, and aiding individuals with evident impairments. Ongoing research in artificial intelligence is increasingly focusing on advancing image

captioning capabilities.

In this realm, locating images, understanding their relationships, and extracting semantic information in natural language are crucial objectives. Notably, efforts in captioning images often employ template-based approaches. These methods require detailing various aspects, including direct or indirect objects, their connections, and associated attributes.

The fundamental framework for these methodologies lies in the encoder-decoder pipeline, comprising two straightforward phases. Initially, characteristics of the image are extracted using Convolutional Neural Networks (CNN) to create encoding through structured embedding vectors. Subsequently, a recurrent neural network (RNN) decoder is commonly utilized for the generation of a linguistic description. The prevalence of neural network-based approaches capable of deducing novel phrases is largely attributed to the strengths of CNN in representation and RNN in temporal modeling.

Advancements in descriptive image technology, as indicated by recent attention and recognition [3], hold the promise of providing a means for visually impaired individuals to perceive the external environment. This field has gained significant importance in the realm of computer vision. Initial methods for generating image descriptions involved the integration of image data through statistical language models and static object class libraries. Notably, the authors propose an approach for automatic geotagging of photos, employing a dependency model to distill information from web pages referencing image locations. L and colleagues have contributed to the field by developing a network-scale n-gram technique. This method aggregates potential terms, synthesizing them to construct sentences that elucidate images from the ground up. The proposed language model is built using the English Gigaword corpus, and the parameters of the hidden Markov model are subsequently derived from these estimations. The resulting picture description is generated by identifying the most probable nouns, verbs, circumstances, and prepositions within the sentences. The overarching objective is to categorize each conceivable region, subjecting it to a prepositional association function, and leveraging a Conditional Random Field (CRF) for predicting image tags. A detector is employed to recognize objects within the image, and 3D image analysis is applied to infer objects, features, and connection points, transforming them into a series of semantic trees.

Following the assimilation of syntax learning to generate written descriptions for semantic trees, the process reversed, transforming these trees back into images. Yagcioglu et al. introduced a query expansion method involving the retrieval of images from a substantial dataset [4]. This approach utilizes the stated distribution in conjunction with the retrieved images, representing one of several indirect methods proposed to address picture description challenges. Once the extended query is generated, suggested descriptions undergo rearrangement by computing the cosine between the separated representation and the expanded query vector. The image's description is then extracted from the closest description in the set. Prior to the era of big data and the widespread adoption of deep learning methods, the effectiveness and broad application of neural networks significantly propelled advancements in the field of photo description, unlocking novel opportunities.

II. Motivation

The field of image captioning, situated at the intersection of computer vision and natural language processing, is dedicated to the task of generating natural language descriptions for images. The fundamental objective of image captioning is to empower computers with the capability to comprehend visual information, enabling them to produce accurate and meaningful descriptions akin to human language. Crafting image captions involves a multifaceted process that encompasses skills such as object detection, scene comprehension, and language modeling. Initially, the system must identify elements within the image, including objects and people, and subsequently leverage this recognition to construct a coherent and precise sentence that aptly describes the depicted scenario. Image captioning holds diverse applications, offering benefits such as assisting the visually impaired, enhancing search outcomes, and refining user experiences across various domains. An illustrative application involves the generation of product descriptions for online shopping platforms, facilitating consumers in visually searching for specific items based on their attributes. Moreover, image captioning plays a crucial role in the advancement of autonomous vehicles, robots, and other artificial intelligence systems that necessitate a profound understanding of and communication with their surroundings. By endowing robots with the ability to precisely describe the visual world, image captioning contributes to the development of more intelligent and adaptable systems, empowering them to comprehend and navigate complex environments effectively. However, despite the recent strides in image captioning research, several challenges persist, such as addressing ambiguity in language and visual data, managing complex scenes featuring multiple objects and actions, and integrating contextual information into the captioning process. Nonetheless, the potential benefits of image captioning are significant [7], and further exploration in this domain is likely to yield meaningful advancements in artificial intelligence and related fields.

III. Main Contributions And Objectives

Illustrate the system's non-functional aspects that are apparent to users and are not directly tied to the system's functional behavior. Non-functional requirements encompass measurable constraints, such as response time (indicating how quickly the system responds to user commands) or accuracy (reflecting the precision of the system's numerical outputs). The primary non-functional requirements for the system include:

- **Usability**: The system is constructed with fully automated operations, eliminating any need for user intervention.

- **Reliability**: The system exhibits increased reliability due to inherent qualities derived from the Python platform. The code is developed using Python, known for its robust reliability.

- **Performance**: The system is being developed using high-level languages and incorporates advanced front-end and back-end technologies. The client system experiences minimal response time.

- **Supportability**: High-level languages are employed in the development of the system, along with advanced front-end and back-end technologies. The client system enjoys a minimal response time.

IV. Proposed Framework

A. Proposed Framework

This paper introduces a novel approach called Reference-based Long Short-Term Memory (R-LSTM) with the primary objective of enhancing descriptive captions for query images by incorporating reference information. During the training phase, the model assigns different weights based on the relationship between images and words. Additionally, the approach aims to maximize agreement scores between captions generated by captioning methods and reference data from adjacent images, addressing the challenge of accurately recognizing an image. In the context of captioning natural scenes, Kiros et al. implemented an encoder-decoder framework known for combining image-text embedding models and multi-modal sentence generation models. This system, resembling language translation, generates a word-by-word output description for a given query image. The textual data is encoded using a specialized Recurrent Neural Network (RNN) called Long Short-Term Memory (LSTM), while visual data is encoded using a deep Convolutional Neural Network (CNN). The visual content is then optimized using a pair-wise ranking loss and fed into an embedding space extended by LSTM hidden states that encode textual data. In this embedding space, a structured content neural language is employed to decode image features, conditioned on the background word's feature vector, allowing for the generation of sentences word by word. Inspired by neural machine translation, Vinyals et al. employed a deep CNN as an encoder for image encoding and an LSTM in RNN as a decoder for decoding purposes, facilitating the generation of descriptions from image features.

B. Benefits of the proposed system

Image captioning proves to be a valuable tool for individuals with visual impairments, providing them with the ability to grasp the content of visuals. Utilizing an AI-powered image caption generator, descriptions of images can be audibly conveyed to those with visual impairments, enhancing their understanding of the environment.

C. Life Cycle Model

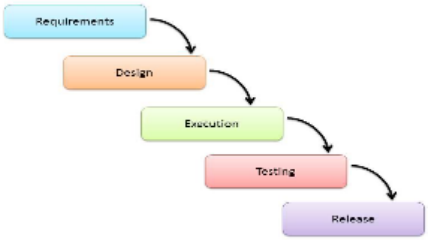

Fig. 1. Life cycle method

In our project, the waterfall model was employed, as it encompasses five distinct steps:

- **Requirements**: During this stage, comprehension of design, functionality, and objectives takes place, and the requirements are documented.
- **Design**: In this stage, the specifications derived from the initial phase's requirements are examined, leading to the creation of the system design. System design, or software architecture, plays a crucial role in defining the hardware, technology, and overall system requirements.
- **Execution**: Guided by the inputs from the system design, the software undergoes a segmentation process into smaller units. This marks the coding phase where the initially outlined requirements are transformed into units, and subsequently integrated to form the final product.
- **Testing**: The product undergoes comprehensive testing across different stages to ensure the absence of errors and the completeness of all requirements. Testing is conducted to mitigate any potential issues that the client might encounter during the software installation process.
- **Release**: After completing the testing phase and confirming the product's completeness with no errors, it is deployed in the customer environment or released into the market

D. DESIGN

1) Architecture: The ultimate framework for generating image captions involves the integration of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). This model takes two inputs: images and corresponding captions. In the RNN layers, each layer receives one word as input, and the model is trained to predict the subsequent word, optimizing itself based on the caption data. Image features are extracted from a pre-trained VGG16 model, stored in a file, and later correlated with the captions. The outputs from both the image features and LSTM layers are amalgamated and input into a decoder model to produce the final captions. The last layer of the decoder has a size equivalent to the vocabulary length. The model

employs the categorical cross-entropy method to predict the probability of each word, and the Adam optimizer is utilized for weight updates during the optimization process.

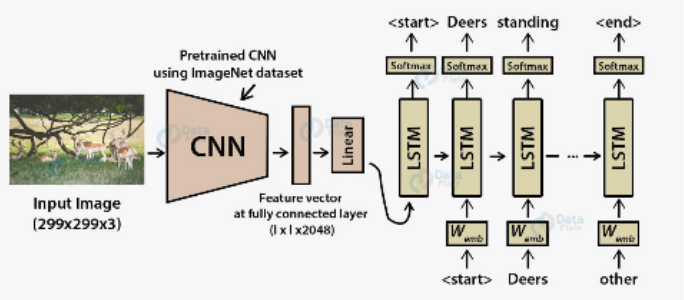

Fig. 2. Model-Image Caption Generator

E. Dataflow Diagram

Data flow diagrams serve as visual depictions of the movement of data within a business information system. They depict the processes related to transferring data from inputs to storage and the generation of reports. Two main categories of these diagrams exist: logical and physical. The logical data flow diagram illustrates how data moves through a system to accomplish particular business functions. In contrast, the physical data flow diagram provides insight into the tangible implementation of the logical data flow.

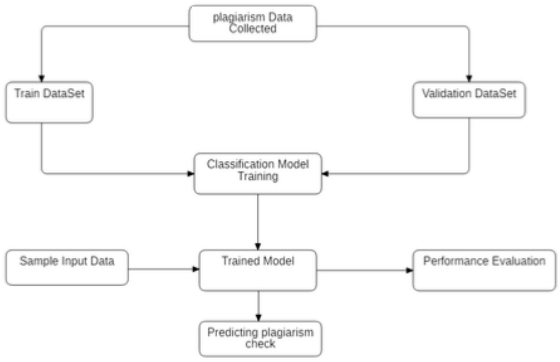

Fig. 3. Data Flow

Design engineering involves the utilization of the Unified Modeling Language (UML), which serves as a standardized language for creating software blueprints. UML is employed as a language for

- Visualizing
- Specifying
- Constructing
- Documenting the artifacts of a software intensive system.

The UML serves as a language that furnishes both vocabulary and rules for combining words within that vocabulary to facilitate communication. A modeling language is characterized by its vocabulary and rules, specifically oriented toward the conceptual and physical representation of a system. Modeling, in turn, contributes to gaining a comprehensive understanding of a system.

V. IMPLEMENTATION

A. Working of Project

1) Dataset: Initially, our model underwent training on the Flickr30k dataset, comprising 31,783 images, each accompanied by five captions. However, challenges arose concerning the model's generalization, stemming from the limited number of training samples and the repetitive nature of the "A man..." template in every caption. To overcome this, we transitioned to the more extensive MSCOCO (2014) training dataset [9], encompassing 82,780 images, each associated with five ground truth captions. For offline evaluation, we employed the Karpathy split3, a non-standardized yet widely utilized split in research, comprising 5,000 images.

2) Syntax Analysis: Syntax identification entails the examination of whether a language adheres to its grammatical rules. Parsing, stemming, and lemmatization represent a few frequently employed techniques within this procedure.

3) Semantic Analysis: Within the realm of Natural Language Processing (NLP), algorithms are applied to grasp the intended meaning. Approaches like Word Sense Disambiguation, Named Entity Recognition, and Natural Language Generation (NLG) are utilized to achieve this goal.

B. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) are widely applied in visual recognition tasks, featuring multiple convolutional layers and fully connected layers. CNNs leverage the 2D structure of images through local connections, tied weights, and pooling techniques, resulting in translation-invariant features. The advantages of employing CNNs include ease of training and a reduced number of parameters compared to other networks with equivalent hidden states. This work utilizes the VGG network, a deep CNN designed for large-scale image recognition, with either 16 or 19 layers, achieving comparable classification error rates for both validation and test sets. The image features extracted by the VGG network play a crucial role in the caption generation process. Long Short-Term Memory (LSTM) [10] is a type of recurrent neural network specifically designed for modeling transient dynamics within a sequence. Traditional RNNs may struggle with learning long-term dynamics due to issues like vanishing or exploding gradients. However, LSTMs address this problem by incorporating a memory cell that retains information over an extended period. Gates control the update timing of the cell state, and the number of connections between the memory cell and gates represents variables.

C. Model deployment

The deployment phase of a model entails putting it into practical use, typically assigned to a data engineer or database administrator after the data scientist has selected a dependable

model and outlined its performance requirements. The deployment process may vary based on the business infrastructure and the specific problem being addressed. The responsibilities of a data engineer encompass implementing, testing, and maintaining infrastructure components to ensure proper data collection, storage, and accessibility. This includes the translation of the final model from high-level programming languages to low-level languages that seamlessly integrate with the production environment.

For assessing the model's performance, data engineers commonly conduct A/B testing, providing insights into how customers interact with a model used for personalized recommendations and its alignment with business goals. Alternatively, in scenarios involving smaller datasets, a database administrator takes on the responsibility of deploying the model into production.

The deployment approach also hinges on whether the earlier stages were executed manually by the data science team using in-house IT infrastructure or automatically through machine learning as a service (MLaaS) products. MLaaS, a cloud platform offering tools for data preprocessing, model training, testing, deployment, and forecasting, introduces variations such as Google Cloud AI, Amazon Machine Learning, and Microsoft's Azure Machine Learning. These MLaaS options differ in the range of provided ML-related tasks, contingent on the level of automation.

Deployment on MLaaS platforms is characterized by automation, and the results can be seamlessly integrated with internal or other cloud-based corporate infrastructures through REST APIs. Additionally, careful consideration should be given to how analytical results are received, whether in real-time or at predefined intervals.

VI. GENERATION OF SENTENCE WITH LSTM

The methodology for generating sentences in a neural network draws inspiration from the encoder-decoder principle utilized in machine translation models. This principle entails employing an encoder to map variable sequences of words in natural language to distributed vectors. Subsequently, these vectors are utilized by a decoder to generate a new sequence of words in the target language. In the training phase, the objective is to optimize the translation process to attain a high probability of generating a natural sentence in the source language. When applied to generating captions for images, the aim is to maximize both the length and quality of the caption produced for a given image.

A. Strategy: matching the problem with the solution

In the early stages of a machine learning project, the emphasis lies in defining strategic objectives. This entails identifying a problem, outlining the project's scope, and planning the development process. A business analyst plays a key role in evaluating the feasibility of a software solution and determining the essential requirements. Simultaneously, a solution architect takes charge of overseeing the development process to guarantee the seamless integration of these requirements into the solution. The solution architect's core responsibility is to verify that the business analyst's established requirements form the fundamental basis for the new solution.

B. Dataset preparation and preprocessing

Data serves as the foundational element for any machine learning project. The second phase of implementing such a project is complex and involves multiple steps, including data collection, selection, preprocessing, and transformation. Each of these steps entails a series of procedures that must be adhered to.

C. Data Collection

In the implementation of a machine learning project, the role of a data analyst is pivotal. They are tasked with identifying pertinent data sources, conducting thorough data collection, interpreting the data, and analyzing the results using statistical techniques. The specific type of data needed varies depending on the nature of the prediction task at hand. Determining the exact amount of data required for a machine learning problem is challenging as each problem possesses its unique characteristics. The selection of attributes used in constructing a predictive model relies on their predictive value. It is generally advisable to gather as much data as possible during the data collection phase, as having more data enhances the chances of accurate model training. Additionally, leveraging publicly available datasets can complement internal data sources, and various platforms like Kaggle, Github, and AWS offer free datasets for analysis.

D. Data visualization

Conveying a substantial amount of information through visual representations enhances comprehension and facilitates analysis. Hence, it is crucial for a data analyst to have proficiency in generating diverse visual aids like slides, diagrams, charts, and templates. For example, utilizing a sales-by-year chart can prove to be an effective means of visually representing sales data.

E. Labeling

Supervised machine learning involves training a predictive model on historical data that contains predefined target answers. This requires providing the algorithm with information about the target attributes or answers to seek within a dataset, a process known as labeling. Labeling datasets can be a laborious and challenging endeavor, especially when dealing with a large volume of records that need labeling for the machine learning model to operate effectively. For instance, if the machine learning algorithm is tasked with classifying different types of bicycles in images, the dataset must be clearly defined and labeled accordingly. One strategy to streamline the labeling process is to involve domain experts to assist in the data labeling.

F. Transfer learning

Transfer learning provides an alternative strategy for dealing with large datasets, allowing the reuse of previously labeled training data. In this approach, insights and knowledge obtained from addressing similar machine learning challenges by other data science teams are leveraged. The process involves a data scientist determining which components of an existing training dataset can be repurposed for a new modeling task. Transfer learning is frequently applied in training neural networks, particularly models utilized for tasks such as image or speech recognition, image segmentation, and human motion modeling.

G. Data selection

Once all the requisite information has been gathered, a data analyst carefully chooses a subset of the data that is pertinent to the specified problem. For instance, if a business is keen on knowing its customers' geographical location, there's no need to include personal details such as cell phone or bank card numbers in the dataset. However, attributes like purchase history would be crucial when constructing a predictive model. This selected subset of data encompasses the attributes that are essential for consideration in the development of a predictive model. In smaller data science teams, it's not uncommon for a data scientist to handle responsibilities ranging from data collection, selection, preprocessing, and transformation to model building and evaluation.

H. Data preprocessing

Preprocessing stands as a pivotal phase in readying data for machine learning. The objective is to convert raw data into a structured and refined format suitable for effective utilization in a machine learning model. This facilitates data scientists in obtaining more precise results from their models. The process encompasses a range of techniques, including data formatting, cleaning, and sampling.

I. Data formatting

In scenarios where data is gathered from diverse sources and individuals, the significance of data formatting escalates. Data scientists initiate the process by standardizing the format of records, ensuring consistent representation of variables for each attribute. This uniformity extends to attributes expressed through numeric ranges. Upholding data consistency is vital for enhancing the accuracy and reliability of machine learning models.

J. Data cleaning

In this phase, a data scientist undertakes various procedures to enhance data quality by eliminating irrelevant information and addressing inconsistencies. Imputation techniques are employed to fill in missing data, and outliers—data points deviating significantly from the rest of the distribution—are detected. If an outlier suggests inaccurate data, the data scientist either eliminates or corrects it. Furthermore, incomplete and irrelevant data objects are removed.

K. Data anonymization

In specific instances, data scientists must eliminate or obfuscate attributes containing confidential information, particularly when handling sensitive data from industries like healthcare or banking.

L. Data sampling

When managing extensive datasets, data analysis demands increased time and computational resources. To tackle this, a data scientist might employ data sampling to select a smaller yet representative subset of the data for model building and execution. This strategy enables quicker and more efficient analysis while still yielding accurate results

M. Data transformation

The final stage of preprocessing entails transforming and amalgamating data to ready it for machine learning or data mining, a practice commonly referred to as feature engineering. This can encompass actions like scaling or normalizing the data and decomposing or aggregating attributes. Such processes guarantee that the data is in a fitting format for analysis and modeling.

N. Scaling

Numeric attributes within data sets may exhibit variations across different ranges, spanning units like millimeters, meters, and kilometers. Scaling involves the transformation of these attributes to a standardized scale, usually ranging between 0 and 1 or 1 and 10, to maintain uniformity in their magnitudes.

O. Decomposition

When confronted with intricate concepts depicted by features, discerning patterns in data can pose challenges. In such scenarios, the application of a method known as decomposition proves beneficial. Through decomposition, higher-level features undergo transformation into lower-level ones, and new features are generated based on existing ones. This technique finds common usage in time series analysis. For example, in predicting the monthly demand for air conditioners, a market researcher might decompose the data representing quarterly demand

P. Aggregation

Aggregation is a methodology that consolidates multiple features into a singular feature encapsulating them all. For instance, customer age data can be aggregated into categories such as 16-20, 21-30, 31-40, and so forth for demographic segmentation. This aids in diminishing the dataset size without compromising valuable information. The preparation and preprocessing of data constitute a gradual and time-consuming process, necessitating the selection of appropriate techniques and iterative adjustments based on the business problem, as well as the quality and quantity of available data.

Q. Dataset splitting

When employing a dataset for machine learning, it is advisable to partition it into three distinct subsets: the training set, test set, and validation set.

R. Training set

A data scientist trains and optimizes a model by utilizing a training set to acquire the essential parameters from the data..

S. Test set

To assess the efficacy of a trained model and its generalization capabilities, a data scientist employs a test set that comprises data not previously utilized for model training. Overfitting occurs when a model is excessively complex and memorizes the training data instead of learning from it. To mitigate this issue, it is crucial to employ distinct datasets for training and testing.

1) Validation set: The validation set plays a pivotal role in fine-tuning a model's hyperparameters, representing the high-level structural settings that cannot be directly learned from the data. These parameters influence a model's complexity and its ability to discern patterns in data. Typically, the dataset is partitioned into a training set (80%) and a test set (20%). The training set undergoes further division, allocating 20% to form a validation set. However, some experts suggest a 66% training and 33% testing split. The size of each subset is contingent on the overall size of the dataset

T. Dataset-splitting

Enhancing a model's potential performance can be achieved by a data scientist through the utilization of a larger amount of training data. Similarly, augmenting the testing data for model evaluation contributes to improving its generalization capability and overall performance

U. Modeling

In this phase, a data scientist trains multiple models to identify which one produces the most accurate predictions.

1) Model training: Upon completing data preprocessing and partitioning the dataset into three subsets, a data scientist initiates the training of various models to identify the one that generates the most accurate predictions. Model training entails providing an algorithm with training data, allowing the algorithm to process it and create a model capable of predicting target values in new data. These target values represent the output of predictive analysis. The choice between supervised and unsupervised learning for model training depends on whether the goal is to forecast specific attributes or group data objects based on similarities.

V. Supervised learning

Supervised learning is a method employed to handle labeled data or data containing target attributes. In this approach, the attributes are already mapped in historical data before training [15] takes place. Through supervised learning, a data scientist can tackle both classification and regression problems.

VII. Results

Fig. 4. Output: a group of giraffes are standing in the grass

Fig. 5. Output: A bird sitting on a branch of tree

Fig. 6. Output: A large white bird flying in the sky

References

[1] Graves, A.; Liwicki, M.; Fernández, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. (May 2009). "A Novel Connectionist System for Unconstrained Handwriting Recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence.

[2] Li, Xiangang; Wu, Xihong (2014-10-15). "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition

Fig. 7. Output: A chair next to a wooden table

[3] Wu, Yonghui; Schuster, Mike; Chen, Zhifeng; Le, Quoc V.; Norouzi, Mohammad; Macherey, Wolfgang; Krikun, Maxim; Cao, Yuan; Gao, Qin (2016-09-26). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

[4] Mayer, H.; Gomez, F.; Wierstra, D.; Nagy, I.; Knoll, A.; Schmidhuber, J. (October 2006). A System for Robotic Heart Surgery that Learns to Tie Knots Using Recurrent Neural Networks. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 543–548.

[5] Rodriguez, Jesus (July 2, 2018). "The Science Behind OpenAI Five that just Produced One of the Greatest Breakthrough in the History of AI". Towards Data Science. Archived from the original on 2019-12-26. Retrieved 2019-01-15.

[6] Li, Xiangang; Wu, Xihong (2014-10-15). "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition

[7] K. Cho, B. Van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empir-ical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014.

[8] Rashtchian C, Young P, Hodosh M, Hockenmaier J. (2010) Collecting image annotations using Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk. Association for Computational Linguistics.

[9] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning

[10] Dewa Made Sri Arsa, I.P.A. Bayupati and Kadek Sastrawan ," Detection of fake news using deep learning CNN–RNN based methods"

[11] Anjali Samad,Bhagyanidhi and Vaibhav Gautam, "An Approach for Rainfall Prediction Using LSTM Neural Network.

[12] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate (2014), arXiv preprint arXiv: 1409.0473.

[13] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neu-ral networks, in: Advances in neural information processing systems, 2014, pp. 3104–3112 [4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Van- houcke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015

[14] H. Liu and P. Singh. ConceptNet - A practical common-sense reasoning toolkit. BT technology journal, 22(4):211–226, 2004.

[15] Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguist 2:67–78