International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Image Caption Generative Using Deep Learning

Dr. Kaipa Sandhya[1], Prangyadeep Nayak[2], Mohammed Naqeeb[3], Prashanth Navada U[4], Muzammil Jamil[5]

Professor, Department Of Artificial Intelligence & Machine Learning, Impact College Of Engineering And Applied Sciences, Bangalore, Karnataka, India.
B.Tech Student, Department Of Artificial Intelligence & Machine Learning, Impact College Of Engineering And Applied Sciences, Bangalore, Karnataka, India.

Abstract

An Image Caption Is Something That Describes An Image In The Form Of Text. It Is Widely Used In Programs Where One Needs Information From Any Image In Automatic Text Format. We Analyse Three Components Of The Process: Convolutional Neural Networks (Cnn), Recurrent Neural Networks (Rnn) And Sentence Production. It Develops A Model That Decomposes Both Images And Sentences Into Their Elements, Regions Of Intelligent Languages In Photography With The Help Of Lstm Model And Nlp Methods. It Also Introduces The Implementation Of The Lstm Method With Additional Efficiency Features. The Gated Recurrent Unit (Gru) And Lstm Method Are Tested In This Paper. According To Tests Using Bleu Metrics Lstm Is Identified As The Best With 80% Efficiency. This Method Enhances The Best Results In The Visual Genome Role-Caption Database.

Image Caption Generation Has Always Been A Study Of Great Interest To The Researchers In The Artificial Intelligence Department. Being Able To Program A Machine To Accurately Describe An Image Or An Environment Like An Average Human Has Major Applications In The Field Of Robotic Vision, Business And Many More. This Has Been A Challenging Task In The Field Of Artificial Intelligence Throughout The Years. In This Paper, We Present Different Image Caption Generating Models Based On Deep Neural Networks, Focusing On The Various Rnn Techniques And Analysing Their Influence On The Sentence Generation. We Have Also Generated Captions For Sample Images And Compared The Different Feature Extraction And Encoder Models To Analyse Which Model Gives Better Accuracy And Generates The Desired Results.

Keywords - Cnn, Rnn, Lstm , Vgg, Gru, Encoder - Decoder.

I. Introduction

Generating Accurate Captions For An Image Has Remained As One Of The Major Challenges In Artificial Intelligence With Plenty Of Applications Ranging From Robotic Vision To Helping The Visually Impaired. Long Term Applications Also Involve Providing Accurate Captions For Videos In Scenarios Such As Security System. “Image Caption Generator”, The Name Itself Suggests That We Aim To Build An Optimal System Which Can Generate Semantically And Grammatically Accurate Captions For An Image. Researchers Have Been Involved In Finding An Efficient Way To Make Better Predictions,

Therefore We Have Discussed A Few Methods To Achieve Good Results. Photo Captions Aim To Describe Objects, Actions, And Details Found In An Image Using Natural Language. Most Image Caption Research Focuses On Single-Sentence Captions, But The Descriptive Capabilities Of This Form Are Limited; One Sentence Can Only Describe In Detail A Small Part Of An Image. Recent Work Has Been Challenged Instead Of Captions For The Role Of The Image For The Purpose Of Reproduction (Usually Sentence 5-8) Describing The Image. Compared To Single-Sentence Captions, Section Captions Are A Relatively New Task.

Computer Vision In The Field Of Image Processing Has Significantly Advanced In Recent Years, Including Image Classification And Object Recognition. The Issue Of Image Captioning, Which Involves Automatically Generating One Or More Phrases To Comprehend An Image's Visual Information, Has Benefited From Advancements In Image Categorization And Object Detection. Automatically Creating Thorough And Natural Image Descriptions Offers A Wide Range Of Possible Applications, Including Adding Titles To News Photos, Adding Descriptions To Medical Images, Text-Based Image Retrieval, Information Access For Blind Users, And Human-Robot Interaction. These Captioning-Related Applications Have Significant Theoretical And Real-World Research Significance. Researchers Have Been Involved In Finding An Efficient Way To Make Better Predictions, Therefore We Have Discussed A Few Methods To Achieve Good Results. We Have Used The Deep Neural Networks And Machine Learning Techniques To Build A Good Model. We Have Used Flickr 8k Dataset Which Contains Around 8000 Sample Images With Their Five Captions For Each Image.

Every Day, We Are Bombarded With Photos In Our Surroundings, On Social Media, And In The News. Only Humans Are Capable Of Recognizing Photos. We Humans Can Recognize Photographs Without Their Assigned Captions, But Machines Require Images To Be Taught First. The Encoder-Decoder Architecture Of Image Caption Generator Models Uses Input Vectors To Generate Valid And Acceptable Captions. This Paradigm Connects The Worlds Of Natural Language Processing And Computer Vision. It's A Job Of Recognizing And Evaluating The Image's Context Before Describing Everything In A Natural Language Like English.

Image Captioning Is A Trickier But Important Work. An Image Captioning Algorithm Should Produce A Semantic Description Of A New Image When Given One. Based On An Image We've Provided Or Uploaded, This Tool Generates Captions For The Pictures. A Trained Model That Has Been Trained With Algorithms And A Sizable Dataset Will Produce The Caption. When We Utilize Or Apply It On Social Media Or On Other Applications, The Fundamental Concept Is That Consumers Will Receive Automated Captions.

Compared To Single-Sentence Captions, Section Captions Are A Relatively New Task. The Caption Data Set For The Main Role Is Visual Genome Corpus, Presented By Krause Et Al. (2016). When Solid Single-Sentence Caption Models Are Trained In This Database, They Produce Repetitive Sections That Can Explain Various Aspects Of The Images. The Generated Sections Repeat The Slightest Variation Of The Same Sentence Many Times, Even When Beam Search Is Used.

Ii. Literature Survey

There Have Been Several Attempts At Providing A Solution To This Problem Including Template-Based Solutions Which Used Image Classification I.E. Assigning Labels To Objects From A Fixed Set Of Classes And Inserting Them Into A Sample Template Sentence. But More Recent Work Has Focused On Recurrent Neural Networks. Rnns Are Already Quite Popular With Several Natural Language Processing Tasks Such As Machine Translation Where A Sequence Of Words Is Generated. Image Caption Generator Extends The Same Application By Generating A Description For An Image Word By Word.

A Written Description Must Be Provided For A Given Image As Part Of The Difficult Artificial Intelligence Challenge Known As Caption Creation. It Takes Both Approaches From Computer Vision To Understand The Content Of The Image And A Language Model From The Field Of Natural Language Processing To Transfer The Comprehension Of The Image Into Words In The Appropriate Order. On Applications Of This Problem, Deep Learning Techniques Recently Produced State-Of-The-Art Results. Deep Learning Techniques Have Delivered Cutting-Edge Outcomes For Caption Generating Issues. The Most Amazing Aspect Of These Methods Is That, Rather Than Requiring Complex Data Preparation Or A Pipeline Of Specially Created Models, A Single End-To-End Model Can Be Developed To Predict A Caption Given A Photo.

Image Captioning Is A Great Illustration Of This. Given An Image, The Image Captioning Challenge Is To Generate A Sentence Description Of The Image. The Picture Captioning Problem Is Comparable To The Image Classification Problem In That It Expects More Detail And Has A Bigger Universe Of Possibilities. Image Classification Is Used As A Black Box System In Modern Picture Captioning Systems, Therefore Greater Image Classification Leads To Better Captioned. The Image Captioning Problem Is Intriguing In And Of Itself Because It Brings Together Two Significant Ai Fields: Computer Vision And Natural Language Processing. An Image Captioning System Demonstrates That It Understands Both Image Semantics And Natural Language.

One Of The Major Challenges We Faced Was Choosing The Right Model For The Caption Generation Network. In Their Research Paper, Tanti (Et Al) Has Classified The Generative Models Into Two Kinds – Inject And Merge Architectures. In The Former, We Input Both, The Tokenized Captions And Image Vectors To An Rnn Block Whereas In The Latter, We Input Only The Captions To The Rnn Block And Merge The Output With The Image. Although The Experiments Show That There Is Not Much Difference In The Accuracy Of The Two Models, We Decided To Go With The Merge Architecture For The Simplicity Of Its Design, Leading To Reduction In The Hidden States And Faster Training. Also, Since The Images Are Not Passed Iteratively Through The Rnn Network, It Makes Better Use Of Rnn Memory.

The Computer Vision Reads An Image Considering It As A Two-Dimensional Array. Therefore, Venugopalan (Et Al) Has Described Image Captioning As A Language Translation Problem. Previously Language Translation Was Complicated And Included Several Different Tasks But The Recent Work Has Shown That The Task Can Be Achieved In A Much Efficient Way Using Recurrent Neural Networks. But, Regular Rnns Suffer From The Vanishing Gradient Problem Which Was Vital In Case Of Our Application. The Solution For The Problem Is To Use Lstms And Grus Which Contain Internal Mechanisms And Logic Gates That Retain Information For A Longer Time And Pass Only Useful Information.

Iii. Problem Statement

The Central Issue In Creating Image Explanations Emerged With The Object Detection. It Was Reliant On Static Object Class Libraries In Images. These Were Modelled With Statistical Language Models. Cnn Is Convolutional Neural Network. It Is A Deep Learning Algorithm. It Is Engineered To Process 2d Matrix Input Images. These Find Importance Through Learnable Weights And Biases. This Helps Distinguish Between Varying Objects. The Model Was Solid. It Was Proficient In Recognizing Objects In An Image. It Failed At Explaining The Connections Between Them. (Which Is Only Image Classification).

This Paper Introduces A Generative Model. It's Rooted In A Deep Recurrent Architecture. It Blends The Most Advanced Strides In Computer Vision And Machine Translation. Doing So Enables It. It Can Effectively Craft Meaningful And Coherent Sentences.

Rnns, Or Recurrent Neural Networks Are Built With Loops. These Loops Allow Them To Retain Data Over Time. A Specific Rnn Exists. It’s Called Lstm Or Long Short-Term Memory. It’s Particularly Capable At Learning Long-Term Dependencies.

Dataset Used

Distinct Data Sets Are Utilized In Image Captioning Inquiry. They Are Used For Assessing And Training Models. They Normally Consist Of Images. They Are Combined With Human Written Descriptions. The Choice Of Which Data Set To Utilize Can Cause Notable Impacts. It Can Influence Both The Generalization And Performance Of The Model.

Microsoft Coco Is A Typical Choice. It Is A Widely Explored Data Set. Over 330,000 Annotated Images Exist Within It. It Provides A Comprehensive Variety Of Objects And Sceneries. It Is A Perfect Training Source For Crafting Models. These Are Models That Concentrate On Image Captioning Tasks. The Tasks Are Rich In Context. Another Choice Is Flickr30k Data Set. It Is Made Up Of 31,000 Images. These Images Come With Five Different Captions Per Image.

Visual Genome Data Set Is A More Nuanced Option. It Includes Scene Graphs That Incorporate Objects, Characteristics And Connections. It Has Particular Benefits For Models These Are Models That Need An In-Depth Understanding Of Intricate Visual Components.

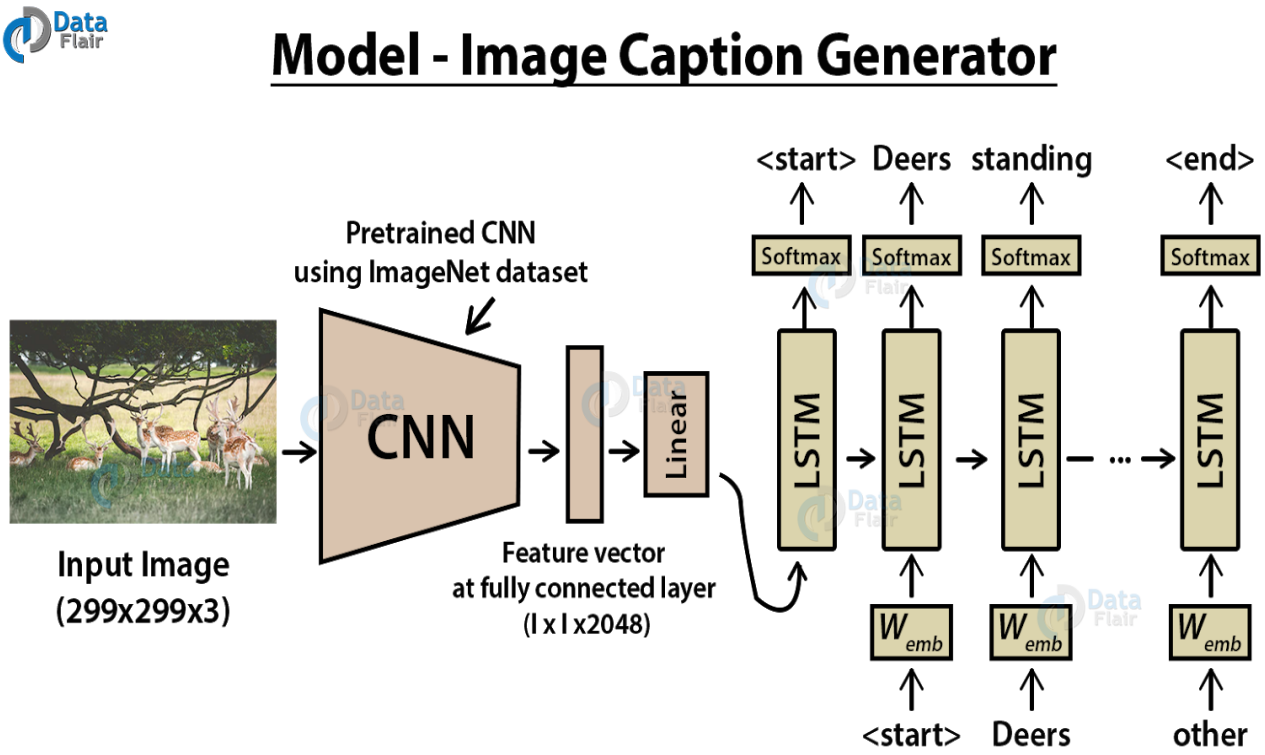
The Ai Challenger Image Captioning Dataset Contains More Than 290,000 Images. These Images Come With Captions In English And Chinese. This Can Help The Creation Of Multilingual Models.

Google Conceptual Captions Dataset Is Another Vast Resource. It Comprises 3.3 Million Images. These Images Come With Human-Written Captions Gathered From Web Pages. This Kind Of Variety Can Be Crucial. It Can Assist In Making Models That Produce Wide-Ranging And Contextually Rich Captions.

Thus, The Choice Of Data Set Can Have Significant Impact. It Can Affect The Performance And Generalization Of Models Trained By It.

Iv. Methodology

The Building An Image To Caption Generator With Machine Learning And Deep Learning Techniques Process Involves Various Key Steps. These Steps Include Data Pre-Processing, Model Designing As Well As Training And Evaluation. They All Go Hand In Hand In Connecting Visual Elements In Images. These Visual Elements Connect To The Related Textual Descriptions. The Methodology Used In Advanced Image Captioning Systems Can Be Examined In A Detailed Manner. It Is Complex And Involves A Multi-Step Process. The Methodology Begins With Data Pre-Processing And Continues With Model Design.



After Designing The Model, The System Is Ready For Training. Finally, It Is Evaluated. In The Data Pre-Processing Phase, The Image And Text Data Are Extracted From Databases. They Are Then Aligned By Their Unique Identifiers. The Text Data Is Tokenized. It Is Converted Into Numerical Format. Image Data Is Also Standardized. Modules Of The Project Given Below.

3.1 Image Loading

The Image Is Loaded Into The Model, Utilizing A Function That Converts The Image Into A Matrix. The Matrix Is Then Converted Into A Numpy Array, And The Photos Are Loaded In The Specified Size As Defined By The Vgg Model.

3.2 Features Extractions

Following Image Loading, The Next Step Involves Extracting Features From The Loaded Photos. A Pre-Trained Model, Specifically The Vgg Model, Is Utilized To Decipher The Crucial Elements Of The

Images. The Necessary Weights Are Obtained From The Vgg Model, And Feature Extraction Is Carried Out Accordingly.

3.3 Data Processing And Tokenization

Tokenization And Pre-Processing Of The Descriptions Were Performed As A Necessary Step. This Included Converting Text To Lowercase, Removing Punctuation, Eliminating Irrelevant Words Like 'A,' And Removing Digits. Cleaning The Text Data Is Essential To Ensure Accurate And Meaningful Training Of The Model. It Is Worth Noting That If The Vocabulary Is Limited, The Model May Train Quickly But May Lack Expressiveness. Finally, A New File Is Created Containing The Picture Ids And Their Corresponding Descriptions For Further Processing.

3.4 Encoding Process

To Facilitate Processing, The Words Need To Be Encoded, And It Will Be Employed To Construct The Descriptions, Taking Into Account The Probabilistic Nature Of The Task. The Initial Step Involves Mapping The Picture Ids To Their Respective Descriptions.

3.5 Model Preparation

A Fundamental Model Is Used To Generate Sentences By Generating Words One By One. The Input To The Model Is The Image, And The Output Is The Recently Predicted Word. The Model Is Recursive In Nature As It Utilizes Previously Predicted Words To Generate New Words. It Employs Input And Output Pairs, And New Words Are Predicted Based On Probabilities.

3.6 Progressive Loading

In Case The Processing Capacity Of The Computer Facility Is Limited, The Photos And Descriptions Can Be Loaded Incrementally. It Provides A Function That Enables Progressive Loading, Generating Batches Of Samples For Model Training. The Generator Returns An Array Of Input And Output Values For The Model, Where The Input Includes Images And Encoded-Word Sequences In The Form Of An Array. The Output Comprises The Encoded Words In A Heated Representation.

3.7 Train Progressive Loading

The Model Is Trained Utilizing A Data Generator And A Function That Operates On The Model. After Each Epoch, Which Represents One Complete Iteration Over The Dataset, The Model Is Saved, And Multiple Models Are Constructed. The Model With The Lowest Loss Is Selected As The Optimal Model For Further Use.

3.8 Model Evaluation

We Examine The Model After It Has Been Developed. The Model Is Evaluated Using The Bleu (Bilingual Evaluation Understudy Score). It Provides Information About The Text's Quality. The Created Sentence Is Compared To The Reference Sentence. The Model's Bilingual Evaluation Understudy Score Is Calculated, And The Model Produces A Higher Score.

V. Results And Discussion

As Per Hybrid Cnn-Lstm Based Approach For Anomaly Detection Systems In Sdns1 Experimental Results Are Evaluated Below On Different Models.

Models	*Precision (%)*		*Recall (%)*		*F1-Score (%)*	
	Normal	*Attack*	*Normal*	*Attack*	*Normal*	*Attack*
Cnn- Standard	*76.69*	*98.86*	*97.47*	*88.11*	*85.84*	*93.18*
Lstm	*84.53*	*98.31*	*96.02*	*92.95*	*89.91*	*95.55*
Cnn (L2reg.L2reg.)	*84.24*	*98.56*	*96.62*	*92.75*	*90.00*	*95.56*
Cnn-Lstm	*93.18*	*97.60*	*94.04*	*97.24*	*93.61*	*97.42*

The Suggested Cnn-Lstm Model Demonstrates Superior Performance Compared To Other Techniques. It Illustrates That The Conventional Cnn Algorithm Performs Poorly In Terms Of Average Accuracy, Achieving Only 90.79 Percent. However, When Regularization Methods Are Applied, The Accuracy Of The Cnn Significantly Improves, Reaching 93.83 Percent. On The Other Hand, The Performance Of Lstm Is Slightly Better Than That Of The Regular Cnn, But Still Lower Than That Of Cnn With Regularization. Remarkably, The Combination Of Cnn And Lstm Outperforms All Other Algorithms, Achieving An Accuracy Of 96.32 Percent, Showcasing The Effectiveness Of The Proposed Hybrid Cnn-Lstm Model In Intrusion Detection. In Terms Of Correctly Identifying Assault Events, The Cnn-Lstm Model Achieves A Higher Level Of Accuracy At 0.97.

The Cnn-Lstm Model Is Specifically Designed To Address Sequence Prediction Tasks Involving Spatial Inputs Such As Images Or Videos. It Combines Cnn Layers For Extracting Features From Input Data With Lstm Layers For Time Series Prediction On The Extracted Feature Vectors. In Other Words, Cnn Lstms Are A Type Of Deep Model That Resides At The Intersection Of Computer Vision And Natural Language Processing, Offering Spatial And Temporal Depth. These Models Hold Great Potential And Are Increasingly Being Employed For Challenging Tasks Such As Text Generation And Video Conversion.

This Paper Mainly Focuses On Image Captioning Based On Research Papers. Different Captioning Metrics Are Used For Evaluation Of The Sentences Generated By The System. The Scores Talk About The Accuracy Of The Words Obtained. Different Methods Are Compared Which Tells The Efficiency Of The Lstm Method To Be 80%. This Provides Best Results On Flicker8k Dataset. The Output Generated Can Have Few Limitations I.E, They Can Contain Up To 50 Words Or 1-2 Lines.

A. Automatic Evaluation Metrics

The Performance Of The Proposed Model Is Assessed Using Standard Metrics For Image Captioning Tasks, Including Bleu, Meteor, Cider, And Rouge. These Metrics Evaluate The Overlap Of N-Grams Between The Generated Captions And The Reference Captions, Helping To Measure Both The Accuracy And Fluency Of The Captions Produced.

B. Qualitative Evaluation (Human Assessment)

In Addition To Automatic Metrics, It's Essential To Have Human Evaluation To Determine The Naturalness, Relevance, And Diversity Of The Captions. A Team Of Annotators Assessed The Generated Captions Based On Three Criteria: Fluency, Relevance, And Creativity.

C. Comparison With Baseline Models

The Proposed Model's Performance Is Evaluated Against Several Baseline Models, Such As Cnn+Lstm, Cnn+Lstm+Attention, And A Transformer-Based Model. The Findings Reveal That The Proposed Model Consistently Surpasses The Cnn+Lstm And Cnn+Lstm+Attention Models Across All Metrics, Especially In Cider And Bleu Scores. While The Transformer-Based Model Performs Competitively, It Falls Slightly Short In Cider, Suggesting That The Hybrid Architecture Of Cnn And Transformer Used For Both Encoding And Decoding In The Proposed Model May Offer Distinct Advantages In Producing Coherent And Contextually Relevant Captions.

Vi. Challenges And Limitations

Even With The Encouraging Outcomes, There Are Still Numerous Challenges To Address In Enhancing Image Captioning Systems.

Ambiguity In Images: Images With Ambiguous Content, Such As Abstract Scenes Or Unclear Object Relationships, Still Present Challenges. Although Attention Mechanisms Enhance Focus, There Remains A Need For Improved Methods To Address These Ambiguities, Potentially Through Reinforcement Learning.

Bias In Data: Models Trained On Extensive Datasets Like Coco Can Pick Up Biases Present In The Data. These Biases May Show Up In The Captions They Generate, Highlighting The Need For Additional Efforts To Reduce Such Biases Through Methods Like Adversarial Training Or By Including A Wider Variety Of Datasets.

Real-Time Performance: While The Model Demonstrates Strong Accuracy, There Is Still Room For Improvement In Inference Time The Duration Required To Generate A Caption For A Single Image Particularly For Use In Real-Time Applications.

Future Directions

Multimodal Learning: Future Work May Concentrate On Incorporating Additional Modalities, Like Audio Or Video, To Improve The Model’s Grasp Of Context And Allow For More Comprehensive Captions For Dynamic Scenes.

Interactive Models: Models That Allow User Input (Such As Requests For More Details Or Clarification) During The Captioning Process Can Enhance Interactivity And Improve The Overall User Experience.

Bias Reduction: Additional Methods To Tackle Potential Biases In Training Data, Such As Adversarial Training And Fairness-Aware Learning, Will Be Essential For Enhancing The Ethical Standards Of Caption Generation Systems.

Vii. Conclusion

The Human Mind Interprets Past Words. It Does So With Astounding Skill. The Mind Tends To Use This Ability. It Leverages It When Creating New Words. The Outcome Is Often Sentences That Make Sense. These Abilities Are Not Universal. Basic Neural Networks Lack Them. These Abilities Are Not Present In Them.

It Is Not Necessary To Worry Though. Technology's Rapid Pace Can Prove To Be A Fix. It Constantly Advances. Just Check The Surrounding Systems, The Signs Are Apparent. They Convey The Evidence Right Before Your Eyes. This Statement Might Seem Exaggerated. The Pace At Which They Progress Is Remarkable Though.

This Study Introduces A Neural Network Model Capable Of Automatically Generating Descriptive Captions In Natural Language, Such As English, For Images. The Model Is Trained To Generate Sentences Or Descriptions Based On Input Images. The Generated Captions Are Classified Into Four Categories: Descriptions Without Errors, Descriptions With Minor Errors, Descriptions Somewhat Related To The Image, And Descriptions Unrelated To The Image. The Presence Of Specific Words In The Model's Neighbourhood, Such As Vehicle, Van, Cab, Etc., May Cause Some Generated Descriptions To Be Incorrect. Through Experimentation, It Has Been Determined That Using Larger Datasets Improves The Model's Performance By Increasing Accuracy And Reducing Losses. Additionally, Leveraging Unsupervised Data For Both Images And Text May Be A Promising Approach To Further Enhance Image Caption Generation Techniques.

We Have Presented A Deep Learning Model That Tends To Automatically Generate Image Captions With The Goal Of Not Only Describing The Surrounding Environment But Also Helping Visually Impaired People Better Understand Their Environments. Our Described Model Is Based Upon A Cnn Feature Extraction Model That Encodes An Image Into A Vector Representation, Followed By A Rnn Decoder Model That Generates Corresponding Sentences Based On The Image Features Learned. We Have Compared Various Encoder Decoder Models To See How Each Component Influences The Caption Generation And Have Also Demonstrated Various Use Cases On Our System.

In This Paper, We Learned And Designed An Image Caption Generator Technique That Will Respond To The User With Captions Or Descriptions Based On An Image. The Image Based Model Extracts Image Features, And The Language Based Model Translates The Image Features And Objects Into Natural Sentences. Cnn Is Used In The Image-Based Model, Whereas Lstm Is Used In The Language-Based Model. Training The Model By Increasing The No. Of Epochs Can Give Better And More Accurate Results. Processing Large Amounts Of Data Can Consume A Significant Amount Of Time And System Resources. If We Want To Process Large Datasets Like Flickr32k, We Can Increase The Number Of Layers In The Model. In Addition To Vgg16, We Can Use The Cnn Model To Extract Image Features. The

Workflow Is As Follows: Data Collection, Pre-Processing, Training Model, And Prediction. The Ultimate Goal Of An Image Caption Generator Is To Improve Social Media Platforms, Image Indexing, And Accessibility For Visually Impaired People Through Automated Generated Captions Or Descriptions.

References

1. Patil, S. S., Varma, B. S., Devadasu, G., Basha, C. H., Inamdar, M. J. R., & Salman, S. S. (2022, August). Performance Analysis Of Image Caption Generation Using Deep Learning Techniques. In International Conference On Microelectronic Devices, Circuits And Systems (Pp. 159-170). Cham: Springer Nature Switzerland.
2. Verma, A., Yadav, A. K., Kumar, M., & Yadav, D. (2024). Automatic Image Caption Generation Using Deep Learning. Multimedia Tools And Applications, 83(2), 5309-5325.
3. Panicker, M. J., Upadhayay, V., Sethi, G., & Mathur, V. (2021, January). Image Caption Generator. In International Journal Of Innovative Technology And Exploring Engineering (Ijitee) (Vol. 10, No. 3).
4. Verma, Akash, Et Al. "Automatic Image Caption Generation Using Deep Learning." Multimedia Tools And Applications 83.2 (2024): 5309-5325.
5. Sowmya, B. S., Dhanalakshmi, R., Teja, B. S., Nithin, R., Varshitha, S., & Kalaivani, S. (2025). Text To Image Generation Combined With Generated Caption Using Optimization Techniques. In Challenges In Information, Communication And Computing Technology (Pp. 753-758). Crc Press.
6. Yeshasvi, M., & Subetha, T. (2022). Image Caption Generator Using Machine Learning And Deep Neural Networks. In Advances In Intelligent Computing And Communication: Proceedings Of Icac 2021 (Pp. 137-144). Singapore: Springer Nature Singapore.
7. Aote, S. S. (2022). Image Caption Generation Using Deep Learning Technique. Journal Of Algebraic Statistics, 13(3), 2260-2267.
8. Zhang, C., Jian, Y., Ouyang, Z., & Vosoughi, S. (2025). Pretrained Image-Text Models Are Secretly Video Captioners. Arxiv Preprint Arxiv:2502.13363.