International Conference on Neural Nexus and Synergy:Innovation in Emerging Technologies
Vol. 1. (2024). PP. 150-162. SNIST

IMAGE CAPTION GENERATOR

1 Gourishetti Nandini, 2 Jilla Varsha Sri, 3 Chennarapu Sarika, 4 SRINATH REDDY CH

1 2 3 Students, SNIST(Sreenidhi institute of science and technology).

4 Asst. professor, SNIST(Sreenidhi institute of science and technology), chsrinathreddy@gmail.com.

Abstract: Image labeling is an interesting job whose goal is to automatically come up with words that describe what's in pictures. Cognitive computing has garnered attention in recent years because to its potential applications in computer vision and natural language processing. Our research aims to construct a complicated Image Caption Generator to assist individuals understand and define themselves. We achieve this using the CNN and LSTM models, two strong neural network architectures. CNN decodes in our system. It looks at the images you give it and pulls out important visual information that you need to understand what's in the pictures. It has been shown that the CNN is very good at finding patterns and things in pictures, which makes it a great part for extracting image features. LSTM, on the other hand, is a processor. It gets the extracted visual features from CNN and turns them into a text that makes sense and explains the picture content. The LSTM is good at this because it can handle data that comes in a certain order and understand how words depend on each other well. By combining CNN and LSTM, our model can easily combine the language knowledge with the visual information from the pictures to make subtitles that are correct and make sense in the given context. Our Image Caption Generator would be able to use this mixed method to include both basic visual details and complex philosophical ideas in the subtitles it creates. After making the captions, we use BLEU Scores to judge the quality of our model. The BLEU measure, which stands for "Bilingual Evaluation Understudy," is often used in NLP tasks to check how close created sentences are to reference sentences. It helps us figure out how well and how quickly our picture annotation system works. In conclusion, our Image Caption Generator is a useful tool for creating natural language descriptions of pictures. Our system can correctly and successfully write subtitles for a wide range of pictures by mixing the power of CNN and LSTM models. This technology has a lot of promise for many uses, such as helping people who are blind or have low vision, making picture search engines better, and making it easier to analyze video material.

Keywords : *Convolution Neural Network (CNN), , Long Short term Memory,BLEU.*

1. INTRODUCTION

There are pictures all around us, on social media, and in the news all the time. People are the only ones who can recognize photos. Image recognition is something that people can do without words, but computers need to be taught how to do it first. Input vectors are used by the encoder-decoder design of picture caption generator models to make subtitles that are correct and relevant. Computer imaging, DL, and NLP are all brought together in this view. Before using a common language like English to describe something, you need to understand and know what the picture is about. Our strategy relies on CNN and LSTM models. The customized software employs CNN to encode and LSTM to decode text and add subtitles to acquire visual features. For example, image captioning can help the blind with text-tospeech by showing realtime information about the scene over a camera feed. It can also improve social medical pleasure by redoing labels for pictures in social feeds and spoken conversations.Helping kids name chemicals is a part of learning the language. Every picture on the internet should have a description. This would make it easier to find real photos and browse through them faster. Images with captions are used in biotechnology, business, the internet, and apps like self-driving cars (which can use them to describe the area around them) and CCTV cameras (which can set off alarms if they see anything suspicious). Simple DL explanations are the focus of this work.

Labeling images requires computer vision and NLP. It is amazing progress in artificial intelligence for a machine to be able to write captions for pictures like a person can. The most important part of this work is showing how the things in the picture are connected in a language that people understand, like English. Within the past, computers have used predefined themes to create written titles for photos. But this method doesn't offer enough variety to make lexically rich text summaries. This flaw is no longer there because neural networks have become more useful. A lot of cuttingedge models use neural networks to make subtitles. They take pictures as input and guess the next word that will be used in the sentence as output.

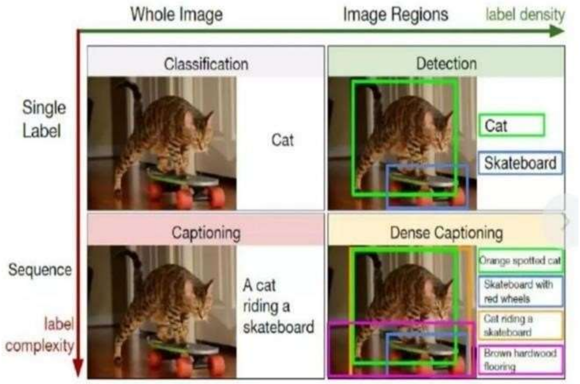


Fig 1 Example Figure

The project's goal is to turn a picture that is fed into it into a writing description. The project's goal is to use DL and NLP to find all the items and characteristics in a picture, figure out how they relate to each other, and then write subtitles that describe each feature. Building an image description generator is the main goal. This way, we can pick a random picture and have our model look at it and come up with some titles.

2. LITERATURE REVIEW

Convolutional Image Captioning:

This is a hard but important job that can be used for virtual helpers, editing tools, picture tracking, and helping disabled people. Its problems come from the fact that there are many different and unclear ways to describe images. Using RNN with LSTM units, picture labeling has come a long way in the past few years. Even though LSTM units are great at remembering relationships and making the disappearing gradient problem less of a problem, they are complicated and naturally work in a certain order over time. New research has shown that neural networks can help with machine translation and conditional picture generation, which is a way to solve this problem. Because of how well they did, we came up with a convolutional picture labeling method in this study. We show that it works on the difficult MSCOCO dataset, where it has the same level of success as the baseline but takes less time to train for each set of parameters. We also do a thorough study and give strong arguments in support of convolutional language creation methods.

Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data:

Recently developed DNN models have shown promise at describing images, but they mostly depend on texts that have both images and sentence subtitles to put things in context. To solve the problem of making descriptions of new things that aren't in paired imagesentence datasets, we present the Deep Compositional Captioner (DCC). Our method does this by using big datasets for object recognition, outside text collections, and sharing information between ideas that are conceptually related. Even though they were taught with big object recognition datasets like ImageNet, current deep caption models can only explain things that are in paired image-sentence texts. On the other hand, our model can write statements that describe new things and how they connect with other things. For example, we show that our model can describe new ideas by testing it on MSCOCO and showing qualitative results on ImageNet pictures of things that don't have paired image-caption data. We also make our method more general by describing things in movie clips. Our results clearly show that DCC is better than other picture and video labeling methods at creating descriptions of new items in their natural setting.

Neural Machine Translation by Jointly Learning to Align and Translate:

New machine translation method is neural machine translation. Neural machine translation uses a single neural network to optimize translation outcomes, unlike statistical machine translation. Encoderdecoder families now include new neural machine translation models. Encoders convert source lines into fixed-length vectors that decoders utilize to translate. We argue in this study that employing a fixed-length vector is hindering this fundamental encoder-decoder design. We aim to enhance it by letting a model automatically (soft)search for source phrase fragments that forecast a target word without having to build them as hard segments. When translating from English to French, this new technique is approximately as fast as the best phrase-based system. Qualitative study confirms the model's delicate linkages with our thoughts.

Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge:

How to automatically describe an image is a major AI challenge that combines computer vision and NLP. We describe a deep recurrent generative model that combines computer vision and machine translation advances. This model generates natural visual descriptions. Learning from the training image, the model improves goal description line likelihood. Tests on diverse data sets reveal that the model is valid and that visual descriptions teach it natural language. We subjectively and statistically test our model to ensure accuracy. Finally, the new COCO dataset was used in a 2015 contest since this task is

so popular. We discuss our baseline adjustments and demonstrate how well it performed in the competition, which we won with a Microsoft
Research team. We provide an open-source TensorFlow application.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention:

We introduce an attention-based approach that automatically describes images. It uses modern machine translation and object recognition research. We discuss training this model deterministically using traditional backpropagation techniques and stochastically by maximizing a variational lower limit. We also use pictures to demonstrate how the model might learn to focus on essential subjects and sequence their words in the output. Attention yielded the greatest results on Flickr8k, Flickr30k, and MS COCO.

3. METHODOLOGY

Our suggested Image Caption Generator uses the power of CNN and Long Short-Term Memory (LSTM) designs to connect language and visual understanding in a smooth way. As the decoder, CNN expertly handles incoming pictures to pull out important visual details needed to understand image information. Image feature extraction is what the CNN does best, and it is famous for how well it can tell the difference between patterns and things. The LSTM, on the other hand, decodes the image content by using the extracted visual features to make words that make sense and describe the image content. The LSTM does its job well because it is good at dealing with sequential data and figuring out how words depend on each other.

Our model can combine visual and verbal information using this combined CNN-LSTM method, which gives us correct subtitles that make sense in the given context. After making captions, we check the quality of our model using BLEU Scores, which are a common way to measure phrase similarity against reference sentences in NLP tasks.

This way of testing makes sure that our picture labeling system works well and correctly. Basically, our Image Caption Generator is a useful tool for explaining pictures in everyday language. It could be used to help people who are blind or have low vision, make image search engines better, and advance the study of multimedia material.

Benefits:

- Our suggested system combines the best parts of CNN and LSTM designs to make it work smoothly with language understanding. This creates subtitles that are very detailed and relevant to the pictures.
- The CNN does a good job of finding patterns and objects in pictures, and the LSTM's ability to work with sequential data and notice word relationships makes sure that statements make sense. This combo makes the process of making subtitles more accurate, so the labels that are made correctly describe what the pictures are about.
- The suggested method is flexible and can make subtitles for a lot of different types of pictures in many different areas. It can handle different kinds of visual material and change to fit different kinds of language and situations, which makes it useful for many things.
- BLEU Scores make sure that the created subtitles are evaluated objectively against reference words, giving a numeric measure of how well the system works. This makes it possible to keep tweaking and improving the model, making sure it works well and is reliable in real life.

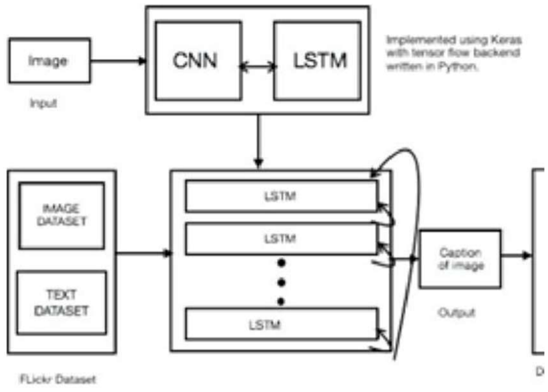


Fig 2 System Architecture

Modules

We have created the following modules in order to carry out the aforementioned project.

- Data exploration: this module will be used to import data into the system.
- Processing: data will be read for processing using this module.
- Splitting data into train & test: The data will be separated into train and test using this module. Building the model – CNN - LSTM
- User signup & login: By using this module, you may register and log in.
- User input: This module provides input for forecasting.

- Prediction: the final forecast is shown

4. IMPLEMENTATION

Algorithms:

CNN:

A Convolutional Neural Network (CNN) is a type of DL program that works really well for jobs that need to recognize and process images. It has many layers, such as fully linked layers, convolutional layers, and pooling layers.

The most important part of a CNN is its

convolutional layers, which are where filters are used on the raw picture to pull out features like lines, colors, and shapes. The convolutional layers' output is then sent to the pooling layers. These layers downsample the feature maps, which means they make the space smaller while keeping the most important data. One or more completely connected layers receive pooling layer output. These layers guess or categorize images.

LSTM:

Long Short-Term Memory RNN. The previous step's result feeds the current step in RNN. Hochreiter and Schmidhuber invented LSTM. It addressed RNNs' long-term dependency, where they couldn't predict the word in their long-term memory but can guess better with additional data. RNN performs poorly as the gap lengthens. LSTM stores data for a long period by default. It predicts, groups, and handles time-series data.

Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) operate well with series data like audio, text, and time series. Sequential data may teach LSTM networks long-term associations. This makes them ideal for language translation, voice recognition, and time series prediction.

The greedy search algorithm is a simple and effective way to decode text. It is used in NLP jobs like picture labeling.

It works in a certain order by guessing one word at a time based on the meanings of the words that came before it.

At each step, the word with the highest conditional chance is picked, which is the best choice in that area. Greedy Search:

Greedy search doesn't look at other word choices or different tracks, so it's easier on the computer than search methods that are more complicated.

It works well for jobs that need to be done quickly and where a slightly less-than-perfect finish is fine.

Beam Search:

Beam search is an intuitive search method that is used to come up with words for things like picture captions.

It goes further than greedy search by looking at more than one possible sequence (beam) at each step.

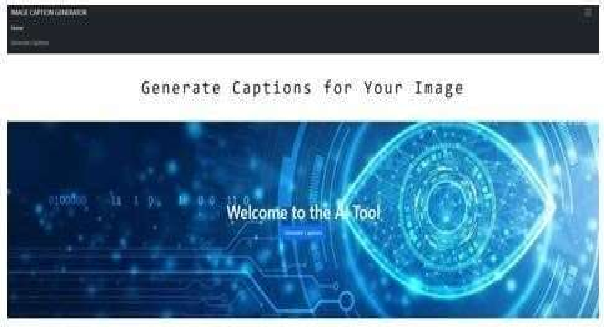
At each step, beam search comes up with possible next words and uses the language model to figure out how likely each one is.

It cuts down the sequences so that only the top beam_index options with the highest total probabilities are kept.

Beam search adds to the sequences that have been kept over and over again until they hit their maximum length or an end code is generated.

The end result is the order that has the best total chance out of all the options that were made.

5. EXPERIMENTAL RESULTS



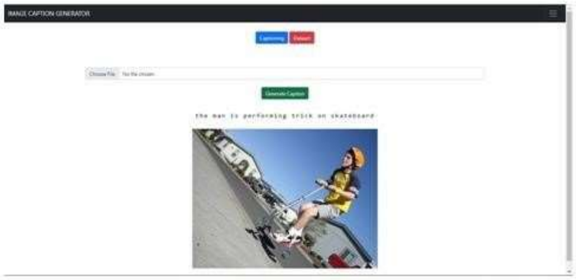
Generate Captions for Your Image

Welcome to the AI Tool

Fig Home page

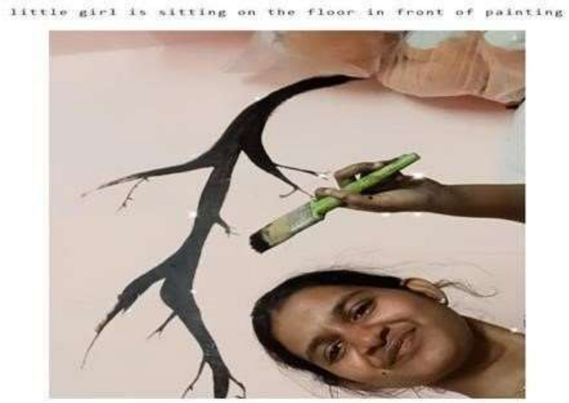
Choose an option

Fig 4 Image upload



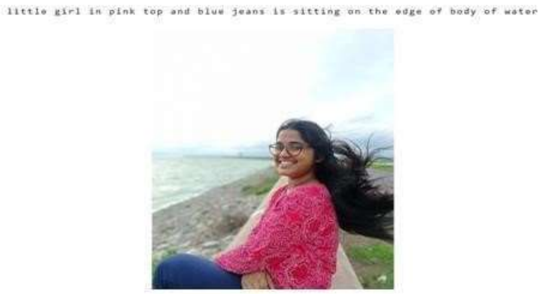
the man is performing trick on skateboard

Fig 5 Beam Search Caption 1



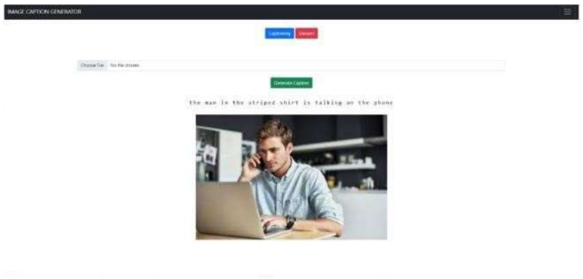
little girl is sitting on the floor in front of painting

Fig 6 Beam search caption



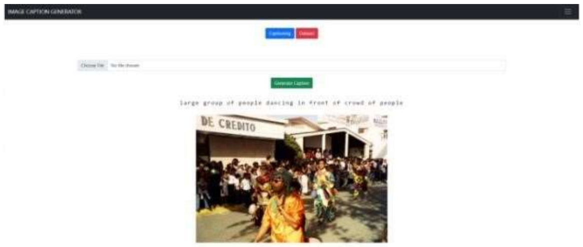
little girl in pink top and blue jeans is sitting on the edge of body of water

Fig 7 Beam search caption



the man in the striped shirt is talking on the phone

Fig 8 Beam search caption



large group of people dancing in front of crowd of people

Fig 8 Beam search caption

6. CONCLUSION

To conclude, the image description project helped computer vision and NLP. By combining DL models with powerful image recognition algorithms, the project has described many images accurately and informatively. The image description generator is very good at understanding and interpreting visual material. It does this by using convolutional neural networks (CNNs) to pull out picture features and recurrent neural networks (RNNs) to generate words. By combining verbal and visual data, the model has been able to come up with subtitles that really get at what the pictures are about and give descriptions that people can understand and enjoy. It can make things easier for people who are blind or have low vision by giving them full explanations of things they can't see. The generator can also be used in photo tracking and search engines, which makes it easier to find specific pictures based on what they're about. In addition, it could be used in robots, social media, and making content.

7. FUTURE SCOPE

Fine-tuning with domain-specific data: Training the model with datasets that are special to a domain can help it do a better job of writing labels for medical images, fashion photos, or sports videos. Adding specific data to the model can help it become more accurate and useful in a certain situation. Multimodal designs: Looking into multimodal systems that can combine text and images well can help make picture captions stronger and more accurate. For a better understanding of visual material, models like Visual Question Answering (VQA) models, which accept both picture and word inputs, can be used. Attention mechanisms: Adding attention mechanisms to the model design can help the creator focus on certain parts of a picture or items when writing descriptions. Attention processes can make it easier for visual and written elements to line up, which can lead to more accurate and relevant subtitles. Better handling of complicated scenes: In the future, it might be helpful to make the model better at understanding complicated scenes, vague ideas, or pictures that aren't clear. This could mean trying new methods like adding more than one text to each picture or using outside sources of information to help the model understand different kinds of visual material better. comments from users and personalization: Letting users give comments on the created titles can help the model get better over time. Adding ways for users to give feedback and using that feedback during training can help create custom comments that are more in line with what each user wants.

REFERENCES

[1]Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5561–5570.

[2]Lisa Anne Hendricks, Subhashini Venu gopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[3]Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (ICLR).Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation.Neuro computing.

[4]Vinyals O., Toshev A., Bengio S., Erhan D.

"Show and tell: Lessons learned from the 2015 mscoco image captioning challenge." IEEE transactions on pattern analysis and machine intelligence. 2017 Apr 1;39(4):652-63.

[5]Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhudinov R., Zemel R., Bengio Y. "Show, attend and tell:Neural image caption generation with visual attention." In International conference on machine learning 2015 Jun 1 (pp. 2048-2057).

[6]Liu C., Mao J., Sha F., Yuille A. L. "Attention Correctness in Neural Image Captioning." In AAAI 2017 Feb 4 (pp. 4176-4182).

[7]You Q., Jin H., Wang Z., Fang C., Luo J. "Image captioning with semantic attention." In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 4651-4659).

[8]Zhao S., Sharma P., Levinboim T., & Soricut R. "Informative Image Captioning with External Sources of Information," arXiv preprint arXiv:1906.08876, 2019.

[9]See A., Liu P. J., & Manning C. D. "Get to the point: Summarization with pointergenerator networks," arXiv preprint arXiv:1704.04368, 2017.

[10]Bahdanau D., Cho K., Bengio Y. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473. 2014 Sep 1.

[11] Holger R. Maier and Graeme C. Dandy,Neural networks for the prediction and forecasting of water resource variables: a review of modelling issues and applica-tions,Environmental Modelling and Software,15,101{124,2000

[12] Avinash N. Bhute and B. B. Meshram,Text Based Approach For Indexing And Retrieval Of Image And Video: A Review,CoRR,abs/1404.1514,2014

[13] Keiron OShea and Ryan Nash,An Introduction to Convolutional Neural Net-works,CoRR,abs 1511.08458,2015

[14] Zachary Chase Lipton and David C. Kale and Charles Elkan and Randall C. Wetzel, Learning to Diagnose with LSTM Recurrent Neural Networks,4th International Confer-ence on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings , 2016

[15] Jurgen Schmidhuber,Deep learning in neural networks: An overview, Neural Networks,85{117,2015

[16] Micah Hodosh and Peter Young and Julia Hockenmaier,Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics,J. Artif. Intell. Res,47,853{899,2013

[17] Tsung-Yi Lin and Michael Maire and Serge J. Belongie and Lubomir D. Bourdev and Ross B. Girshick and James Hays and Pietro Perona and Deva Ramanan and Piotr Dollar and C. Lawrence Zitnick,Microsoft COCO: Common Objects in Context,Computing Re-search Repository (CoRR),abs/1405.0312,2014

[18] Karen Simonyan and Andrew Zisserman,Very Deep Convolutional Networks for Large-Scale Image Recognition,Computer Science - Computer Vision and Pattern Recogni-tion,2014

[19] Polina Kuznetsova and Vicente Ordonez and Alexander C. Berg and Tamara L. Berg and Yejin Choi,Collective Generation of Natural Image Descriptions,359{368,The Association for Computer Linguistics,2012

[20] Siming Li and Girish Kulkarni and Tamara L. Berg and Alexander C. Berg and Yejin Choi,Composing Simple Image Descriptions using Web-scale N-grams,Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Port-land, Oregon, USA,220{228,ACL,2011