IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Exploring Image-Text Alignment with Flickr Data: A Deep Learning Approach

Chiriki Usha [1], Geetha Arsha Surisetty [2], Kodukula Satya Gopal [3], Sumith Kumar Poddar [4], Mathurthi Pavani [5]

[1] **Assistant Professor**, Department of Computer Science & Engineering – AI & ML, Dadi Institute of Engineering and Technology , NH-16, Anakapalle, Visakhapatnam-531002,A.P

[2,3,4,5] **B. Tech Students**, Department of Computer Science & Engineering - AI & ML, Dadi Institute of Engineering and Technology, NH-16, Anakapalle, Visakhapatnam-531002,A.P

Abstract

The Image Caption Generator project aims to bridge the gap between visual data & NL by automatically generating descriptive captions for images. The model combines the strengths of CNNs and LSTM networks. Specifically, VGG16, a pre-trained CNN model, is used to extract deep features from images, which are then passed to an LSTM network to produce captions. Model's performance is evaluated using the BLEU score to assess how well the generated captions align with human-written ones. The Flickr 8k dataset, containing 8 thousand images each paired with five captions, serves as the primary dataset for training and testing. Implemented in Python, Jupyter Notebook, and Google Colab, the model is flexible and scalable, making it suitable for various real-world usage like accessibility and content management.

Keywords

Image Captioning, CNN, LSTM, VGG16, BLEU Score, Flickr 8k Dataset, Deep Learning.

Introduction

Creating meaningful captions for images is a key challenge that brings together computer vision and NLP. Traditional methods for picture captioning often fall short when it comes to capturing both the visual and contextual information necessary to generate natural- sounding descriptions. With advancements in deep learning, we now have more sophisticated tools, such as CNNs and RNNs, to address these challenges.

This project focuses on developing a deep learning model using a combination of VGG16 for feature extraction & LSTM for generating sequences of text. The goal is to achieve accurate image-text alignment and provide a scalable solution that can be applied in various fields, including social media, accessibility, and AI-based image indexing.

Dataset and Data Description

The dataset used for this project is the Flickr 8k dataset, which have 8,000 samples, each annotated with 5 descriptive captions. This dataset is ideal to teach models that need to learn the mapping between visual features and textual descriptions.

Dataset Overview:

- Image Data: 8,000 images in JPG format.
- Captions: Each image is paired with 5 unique captions, providing a diverse set of descriptions.

The dataset is categorised into training & testing sets. During training, model learns to predict captions based on the image data, and performance is evaluated using the BLEU score to measure how closely the generated captions match the human-written captions.

Motivation / Literature Survey

The increasing demand for automated solutions that can generate natural language descriptions for visual data has led to significant advancements in image captioning. Traditional methods have relied on manually crafted rules or simple machine learning algorithms, which struggle with the complexity of modern visual datasets. Recent research has shown that deep learning models, especially CNNs and LSTMs, are particularly effective for this task.

Several research have explored the use of CNNs for feature extraction and LSTMs for language modeling. CNNs, such as VGG16, are excellent at extracting high-level features from images, while LSTMs are well-suited for sequence prediction tasks. This combination has been shown to create state-of-the-art results in image captioning tasks.

Algorithms and Implementation

Model Architecture

- VGG16 (CNN): Used to extract deep visual features from images.
- LSTM: A recurrent neural network used for generating sequential descriptions (captions). - Evaluation Metric (BLEU Score): It performes how closely the generated captions match, ground truth captions by comparing n-grams.

Implementation Details

- Libraries Used: Python, TensorFlow, Keras, Jupyter Notebook, Google Colab.
- Feature Extraction: The VGG16 model, pre-trained on ImageNet, extracts feature vector from each image.
- Caption Generation: The extracted features are passed to the LSTM, which generates a caption one word at a time.
- Evaluation: The BLEU score is calculated to evaluate the quality of the Output captions.

Architecture / Workflow

Step 1: Data Preprocessing

- Image Preprocessing: Resize images to 224x224 pixels and normalize them.
- Text Preprocessing: Tokenize the captions and convert them into sequences for training.

Step 2: Feature Extraction (VGG16)

- Extract a fixed-size feature vector for each sample using the pre-processed VGG16 model.

Step 3: Caption Generation (LSTM)

- The LSTM takes the extracted image features as input and produces a caption word by word.

Step 4: Model Training

- The model is trained using the Flickr 8k dataset, where the LSTM learns to predict the next word in a caption sequence based on the previous words and the image features.

Step 5: Evaluation (BLEU Score)

- The BLEU score is used to evaluate the correlation between the generated captions and the actual captions.

Future Scope

There are several avenues to enhance the working of image captioning models. Future work can explore:

- Attention Mechanisms: Implementing attention performance to allow the model to focus on different parts of the samples while generating captions.

- Transformers: Utilizing transformer architectures, such as BERT or GPT, to improve the fluency and accuracy of the produced captions.

- Larger Datasets: Training the model on larger datasets like MSCOCO to improve its generalization to new images.

Conclusion

This project demonstrates a deep learning approach to image caption generation using a combination of VGG16 and LSTM. By leveraging the Flickr 8k dataset and evaluating the model with the BLEU score, we have developed a system capable of generating descriptive captions for images. The approach shows promise for a variety of applications, including accessibility for visually impaired users, social media content management, and AI-powered image indexing.

References

[1] Abhaya A and Alon L. 2008.Meteor, m-Belaunde-Ter: Eval metrics for high- correlation with Human rank of MT on o/p. In Proceed of the 3rd W.S.M., ACL. ,115–118.

[2] Ahmet Akerand R Gaizauskas. 2010. Generating img dcp using dependency rel.patterns. In Proceedings of the 48th AMoACGL,1250–1258.

[3] Peter A, Bansuri F, Mark J son, and S Gould. 2016. Spice: Semantic prop img Capt eval. In ECEVS,382–398.

[4] P Anderson, Chris B D.Teney, Mark J son, S Gould, & Lei Zhang. 2017. Bottom-up and top-down attend for img capt and vqa. ArXiv preprint arXiv:1707.07998 (2017).

[5] Jyoti A, Aditya D, and G Schwing. 2018. Conv image captioning. In Proceedings of The IEEE CoCV -PR.5561–5570.