EasyChair Preprint

№ 10803

NLP Metrics Suitable for Various Categories of Image Captions

B Kezia Rani, Prajwal Reddy Dornala and Sravani Golla

August 30, 2023

NLP Metrics suitable for various categories of Image Captions

1st Dr. B Kezia Rani
Department of Infromation Technology
(Associate Professor)
Vasavi College of Engineering
(Osmania University)
Hyderabad, India
Keziarani2020@gmail.com

2nd D. Prajwal Reddy
Department of Infromation Technology
(Student)
Vasavi College of Engineering
(Osmania University)
Hyderabad, India
dornalaprajwalreddy@gmail.com

3rd G. Sravani
Department of Infromation Technology
(Student)
Vasavi College of Engineering
(Osmania University)
Hyderabad, India
sravanigollagsy@gmail.com

***Abstract*—**

The aim of this project is to explore and identify suitable natural language processing (NLP) metrics for different styles and types of image captions. Image captions play an important role in providing additional context and information about images, and the effectiveness of the caption depends on various factors such as style, type, and category. The project will utilize a diverse set of image caption datasets from different domains, such as social media, news, scientific research, and other sources. Various NLP metrics, including sentiment analysis, readability, coherence, and topic modeling, will be applied to these datasets to evaluate their effectiveness in capturing the nuances of different caption styles and types. The project will analyze the effectiveness of each metric in capturing the nuances of different caption styles and types, such as humorous, informative, and poetic captions, as well as captions for different image categories, such as nature, sports, and food. The project will also investigate the impact of other factors, such as the length, structure, and content of the captions, on the applicability of NLP metrics. The findings of this project will contribute to the development of more accurate NLP models for analyzing different types of image captions, which will assist in improving the user experience of various image-based applications. Moreover, it will provide valuable insights into the relationship between the content of the image and the quality of the caption, enabling better captioning algorithms in the future.

Keywords*— *NLP, CNN, SPICE, BLEU, METEOR, ROGUE.

EASE OF USE

Abbreviations and Acronyms

1. NLP - Natural Language Processing

2. VGG - Visual Geometry Group

3. LSTM - Long Short-Term Memory

4. CNN - Convolutional Neural Networks

5. CIDEr - Consensus-based Image De- scription Evaluation

6. GAN - Generative Adversarial Network

7. BLEU - Bleu bilingual evaluation understudy

8. COCO - Common Objects in Context

9. R-CNN - Region-based Convolutional Neural Network

10. RNN - Recurrent Neural Networks

11. SPICE - Semantic Propositional Image Caption Evaluation

12. MTEOR - Metric for Evaluation for Translation with Explicit Ordering

I. INTRODUCTION

Natural Language Processing (NLP) techniques have become increasingly important in various applications, such as machine translation, sentiment analysis, and text summarization. NLP metrics are used to evaluate the performance of these techniques and to compare different models. However, NLP metrics are not one-size-fits-all and may perform differently depending on the characteristics of the text or captions.

Captions are often used to describe images or videos and convey information in various contexts, such as social media, news articles, and scientific publications. Captions can vary widely in their style, semantics, category, meaning, summary, and other aspects, depending on the context and purpose of the visual content. Therefore, it is important to identify NLP metrics that are suitable for different categories of captions or text, to ensure accurate and effective caption generation, and evaluation.

The findings of this project can also provide insights into the strengths and limitations of existing NLP techniques and metrics and can guide the development of more effective techniques for caption generation and evaluation.

proposed method can be utilized to aid farmers and experts in detecting and identifying plant diseases in a timely and accurate manner, which can significantly reduce the economic and environmental impact of plant diseases.

II. LITERATURE SURVEY

Natural Language Processing of Text-Based Metrics for Image Captioning by Sudhakar Sengan –

A rapidly increasing interest has been focused on natural language processing and computer vision study by automatically producing descriptive sentences for images in present trends. Image processing is a success factor, especially involving somaticized image knowledge and creating reliable, appropriately organized explanation phrases. In this paper, a dynamic program that uses the VGG16 platform to develop image descriptions and a long short-term memory (LSTM) to contain appropriate words with generated text keywords successfully is proposed. This paper determines the NLP method's usefulness using Flickr8 K-statistical models and demonstrates that their model delivers accurate performance compared to the Bleu metric. The Bleu metric is an automated system for measuring a machine translation's performance by monitoring the effectiveness of text transformed from natural to other languages. The results have been justified with images that satisfy the NLP algorithm.

Learning-based Composite Metrics for Improved Caption Evaluation by Naeha Sharif, Lyndon White, Mohammed Bennamoun –

In our approach, we use scores conferred by a set of existing metrics as an input to a multi-layer feed-forward neural network. We adopt a training criteria based on a simple question: is the caption machine or human generated? Our trained classifier sets a boundary between good and bad quality captions, thus classifying them as human or machine produced. Furthermore, we obtain a continuous output score by using the class-probability, which can be considered as some "measure of believability" that the candidate caption is human generated. Framing our learning problem as a classification task allows us to create binary training data using the human generated captions and machine generated captions as positive and negative training examples respectively.

Our proposed framework first extracts a set of numeric features using the candidate "C" and the reference sentences "S". The extracted feature vector is then fed as an input to our multi-layer neural network. Each entity of the feature vector corresponds to the score generated by one of the four measures: METEOR, CIDEr,WMD1 , SPICE.

Image Caption Metrics of Methods by Omkar Sargar –

Image Captioning is one of the emerging topics of research in the field of AI. It uses a combination of Computer Vision (CV) and Natural Language Processing (NLP) to derive features from the image, use this information to identify objects, actions, their relationships, and generate a description for the image. It is most important concept in artificial intelligence applied in the fields like aid to the blind, self-driving cars, and many more. This paper we demonstrate a concise state of art image captioning and its method for caption generation using deep learning concepts. We also determine the approach for image caption generation using Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN) model in deep learning framework. Using this approach system intelligent enough to create sentences for images. It uses the encoder-decoder architecture, where CNN is used for image vector generation and LSTM is used for the generation of a logical sentence using the NLP concepts. Finally, we evaluate the proposed system experimental analysis with numerous existing systems and show the effeteness of system.

Semantic Propositional Image Caption Evaluation P.Anderson, B.Fernando, M.Johnson and S.Gould –

The Semantic Propositional Image Caption Evaluation (SPICE) is a metric that evaluates the quality of image captions based on their semantic content. Unlike other metrics that primarily focus on measuring the similarity between the generated and reference captions, SPICE considers the semantic content of the captions and how well they convey the meaning of the associated image.

SPICE represents the meaning of the caption and the image using a scene graph, which is a graph-based representation of the objects, attributes, and relationships present in the image. The caption is then transformed into a set of semantic propositions, which are compared to the scene graph using a set of semantic similarity measures. The final score is a weighted combination of these measures.

The SPICE metric has been shown to be effective at evaluating the semantic content of image captions, and is able to capture aspects of meaning that are not captured by other metrics. It has been used to evaluate the performance of various image captioning models on benchmark datasets, and has become a widely used metric in the field of image captioning.

Overall, the SPICE metric provides a valuable tool for evaluating the semantic content of image captions and improving the performance of image captioning systems.

Image captioning: transforming objects into words by

S. Herdade, A. Kappeler, K. Boakye, and J. Soares –

In the paper "Image captioning: transforming objects into words" by S. Herdade, A. Kappeler, K. Boakye, and J. Soares, the authors present a novel approach to image captioning that transforms objects in an image into words in a sentence.

The proposed model consists of two main components: an object detection network that detects the objects present in an image, and a language model that generates a caption based on the detected objects. The object detection network is based on the Faster R-CNN architecture and is used to detect objects in the image along with their bounding boxes. The language model is based on a variant of the transformer architecture and generates a caption by attending to the detected objects.

The authors also propose a novel attention mechanism that combines bottom-up and top-down attention. The bottom-up attention mechanism attends to the detected objects in the image, while the top-down attention mechanism attends to the language model and helps to guide the generation of the caption.

The proposed model was evaluated on the MS COCO dataset and achieved state-of-the-art performance on several metrics, including BLEU-4, METEOR, and CIDEr. The authors also conducted several experiments to analyze the contribution of different components of the model, and showed that the proposed attention mechanism and object detection network were critical for achieving high performance.

Overall, the proposed approach demonstrates the effectiveness of transforming objects in an image into words in a sentence for image captioning, and the importance of combining bottom-up and top-down attention mechanisms in this task. The results suggest that this approach could lead to further improvements in image captioning performance.

Microsoft COCO Captions: Data Collection and Evaluation Server by -

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam Saurabh Gupta, Piotr Dollar, C. Lawrence Zitnick

The paper "Microsoft COCO Captions: Data Collection and Evaluation Server" by Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick presents the Microsoft COCO dataset, which is a large-scale image captioning dataset.

The authors collected the dataset by selecting images from the Microsoft Common Objects in Context (COCO) dataset and asking human annotators to provide captions for each image. The resulting dataset contains over 330,000 images and more than 5 million captions. The captions were annotated for accuracy, diversity, and fluency, and were evaluated using several metrics, including BLEU, METEOR, and CIDEr.

The authors also present an evaluation server that allows researchers to evaluate their image captioning models on the Microsoft COCO dataset using the same metrics used for the dataset creation. The server also includes a leaderboard that tracks the performance of different image captioning models on the dataset.

The Microsoft COCO dataset has become a widely used benchmark for image captioning research and has led to significant advances in the field. The authors note that the dataset is not only useful for image captioning, but also for other computer vision tasks such as object detection and segmentation.

Overall, the Microsoft COCO dataset and evaluation server provide valuable tools for evaluating and advancing the state of the art in image captioning and related computer vision tasks.

Various NLP metrics that can be used for image captioning:

1. BLEU: BLEU (Bilingual Evaluation Understudy) is a commonly used metric for evaluating machine translation systems, but it has also been applied to image captioning. The metric compares the generated captions to a set of reference captions and computes the n-gram overlap between them. The BLEU score ranges from 0 to 1, with higher scores indicating better performance. However, BLEU has been criticized for not capturing the semantic content of the generated captions, and for being overly sensitive to minor variations in the wording of the reference captions.

2. METEOR: METEOR (Metric for Evaluation of Translation with Explicit Ordering) is another commonly used metric for image captioning. It is similar to BLEU, but it also takes into account semantic similarities between the generated captions and the reference captions. METEOR uses a combination of unigram matching, stemming, and synonymy to compute the similarity between the generated captions and the reference captions. The metric has been shown to correlate well with human judgments of caption quality.

3. ROUGE: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a family of metrics that are commonly used for text summarization, but can also be used for image captioning. They compute the overlap between the generated caption and the reference captions at various levels of granularity (e.g., words, n-grams, and sentences). The ROUGE metrics have been shown to correlate well with human judgments of caption quality, but they have also been criticized for not capturing the diversity of the generated captions.

4. CIDEr: CIDEr (Consensus-based Image Description Evaluation) is a metric that is designed to capture the diversity of the generated captions. It computes the similarity between the generated captions and the reference captions at the level of individual words, and then combines these similarities to produce an overall score. CIDEr has been shown to outperform other metrics on certain image captioning datasets, and it has been used in a number of state-of-the-art image captioning models.

5. SPICE: SPICE (Semantic Propositional Image Caption Evaluation) is a metric that evaluates the semantic content of the generated captions. It uses semantic parse trees to compare the generated captions to the reference captions, and it can capture a wide range of semantic relationships between words. SPICE has been shown to correlate well with human judgments of caption quality, and it has been used in a number of recent image captioning models.

6. GLEU: GLEU (Google-BLEU) is a metric that is similar to BLEU, but it uses a different smoothing method that is based on Google's n-gram models. It has been shown to outperform BLEU on certain image captioning datasets, and it has been used in a number of recent state-of-the-art image captioning models.

7. WMD: WMD (Word Mover's Distance) is a metric that measures the semantic similarity between two sets of words. It has been used in image captioning to compare the generated captions to the reference captions at a semantic level. WMD has been shown to correlate well with human judgments of caption quality, but it can be computationally expensive to compute.

These are just a few examples of the many NLP metrics that have been used in image captioning research. Each metric has its own strengths and weaknesses, and the choice of metric depends on the specific research question, the dataset, and the characteristics of the image captioning model being evaluated. It is common to evaluate models using multiple metrics to get a more comprehensive understanding of their performance on the specific research question, the dataset, and the characteristics of the image captioning model being evaluated.

Importance of large-scale datasets for advancing the state-of-the-art in image recognition by Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.

The paper "ImageNet: A Large-Scale Hierarchical Image Database" was presents on Computer Vision and Pattern Recognition. The paper presents the ImageNet database, which is a large-scale dataset of over 14 million images, each annotated with hierarchical labels. The database was created to provide a benchmark dataset for computer vision research and has been used extensively in a variety of tasks such as object recognition, detection, and segmentation.

The paper describes the process of creating the ImageNet database. The authors collected the images from the Internet and used a combination of automatic and manual techniques to remove duplicates and low-quality images. They then used crowdsourcing to annotate the images with hierarchical labels. The labels are organized in a tree structure, with more general categories at the top and more specific categories at the bottom. For example, the top-level categories include things like "animal" and "vehicle", while the bottom-level categories include things like "dog" and "car".

The authors also discuss the use of the dataset in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which has been held annually since 2010. The challenge involves training algorithms on a subset of the ImageNet dataset and then evaluating their performance on a separate test set. The challenge has become a widely recognized benchmark for evaluating the performance of image recognition algorithms, and has led to significant advances in the field.

The paper concludes with a discussion of some of the challenges associated with creating and using large-scale datasets like ImageNet. One challenge is the cost and effort required to collect and annotate such a large dataset. Another challenge is the potential bias in the dataset, which can affect the performance of algorithms trained on the dataset. The authors suggest that these challenges can be addressed by continued efforts to improve the quality and diversity of the dataset, and by developing techniques to mitigate the effects of bias.

Overall, the paper presents the ImageNet database as a valuable resource for researchers in the field of computer vision, and highlights the importance of large-scale datasets for advancing the state-of-the-art in image recognition. The paper has been cited over 87,000 times according to Google Scholar, indicating its significant impact on the field.

Aligning linguistic words and visual semantic units for image captioning by Guo, L., Liu, J., Tang, J., Li, J., Luo, W., Lu, H

The paper "Aligning Linguistic Words and Visual Semantic Units for Image Captioning" proposes a novel approach to image captioning that aligns linguistic words with visual semantic units to generate more accurate and coherent captions.

The paper begins by discussing the current state of the art in image captioning, which typically involves training a neural network to generate captions based on visual features extracted from the image. However, these approaches often produce captions that are generic and lack specific details, which can lead to inaccurate and uninformative captions.

The proposed approach seeks to address this issue by aligning linguistic words with visual semantic units, which are learned representations of visual concepts such as objects, attributes, and relations. The authors use a graph-based approach to model the relationship between words and visual semantic units, and then use a reinforcement learning algorithm to train the model to generate captions that align well with the visual semantic units.

The paper presents experimental results on two benchmark datasets, COCO and Flickr30k, demonstrating that the proposed approach outperforms state-of-the-art methods in terms of both quantitative metrics and human evaluation. The authors also conduct ablation studies to analyze the contribution of different components of the model, and provide qualitative analysis of the generated captions.

Overall, the paper proposes a novel approach to image captioning that aligns linguistic words with visual semantic units, and demonstrates its effectiveness through experimental results on benchmark datasets. The paper has significant implications for the field of computer vision, as accurate and informative image captions are essential for a wide range of applications such as image retrieval, assistive technology, and social media analysis.

Bottom-up and top-down attention for image captioning and visual question answering by Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.

The paper "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering" presents a new approach to image captioning and visual question answering that combines bottom-up and top-down attention mechanisms. The paper was presented at the IEEE Conference on Computer Vision and Pattern Recognition in 2018.

The proposed model consists of two main components: a bottom-up attention mechanism and a top-down attention mechanism. The bottom-up attention mechanism is designed to identify salient regions in the input image, while the top-down attention mechanism is used to focus on specific features of the image that are relevant to the given task.

The authors use a convolutional neural network (CNN) to extract visual features from the input image, which are then used to compute the bottom-up attention maps. The top-down attention mechanism is implemented using a recurrent neural network (RNN) that generates captions or answers to questions based on the input image and the attention maps.

To evaluate the proposed approach, the authors conducted experiments on two benchmark datasets: the COCO dataset for image captioning and the VQA v2.0 dataset for visual question answering. The results show that the proposed approach outperforms state-of-the-art methods in terms of both quantitative metrics and human evaluation.

The paper also includes ablation studies to analyze the contribution of different components of the model, and provides qualitative analysis of the generated captions and answers. The authors show that the proposed approach is able to generate more accurate and informative captions and answers than previous methods.

Overall, the paper presents a novel approach to image captioning and visual question answering that combines bottom-up and top-down attention mechanisms, and demonstrates its effectiveness through experimental results on benchmark datasets. The proposed approach has significant implications for a wide range of applications such as autonomous vehicles, robotics, and augmented reality, where accurate and informative image captions and answers to questions are essential.

III. PROPOSED INVESTIGATION

Proposed work for a project that aims to find suitable NLP metrics for each category of statements that differ based on style, semantics, sentiment, form, type, etc. may include the following steps:

1. Data collection: Collect a diverse range of textual data that covers various categories of statements. This could include news articles, social media posts, scientific texts, literary texts, product reviews, etc. The data should be annotated with labels that capture different aspects such as sentiment, style, form, etc.

2. Literature review: Conduct a comprehensive literature review to identify existing NLP metrics and evaluation techniques that are suitable for the different categories of statements. This should include an analysis of the strengths and weaknesses of different metrics in capturing the various aspects of the text.

3. Feature extraction: Extract relevant features from the collected textual data that capture different aspects such as style, form, sentiment, etc. These features could include syntactic and semantic features, sentiment scores, readability measures, etc.

4. Metric selection: Select a set of NLP metrics that are suitable for each category of statement based on the analysis of the literature review and feature extraction. The selected metrics should be able to capture the relevant aspects of the text for each category.

5. Empirical evaluation: Evaluate the selected metrics on the collected textual data for each category of statement. This evaluation should be conducted using appropriate statistical methods, such as regression analysis or correlation analysis.

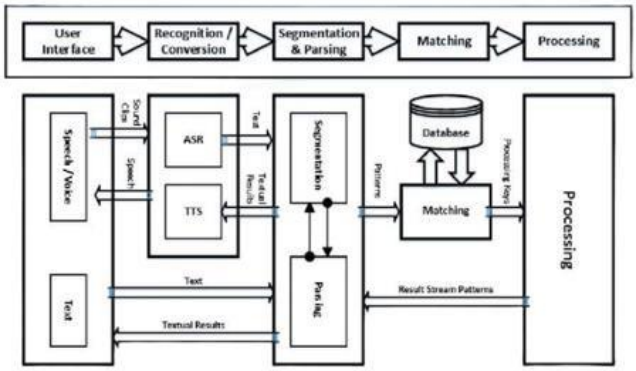
6. Model optimization: Optimize NLP models for each category of statement based on the selected metrics. This may involve fine-tuning existing models or developing new models that are specifically optimized for each category.

7. Model comparison: Compare the performance of the optimized NLP models across the different categories of statements using the selected metrics. This comparison should provide insights into which models perform best for different categories and which metrics are most appropriate for each category.

8. Analysis and reporting: Analyze the results of the empirical evaluation and model comparison and report the findings. This may include insights into which NLP metrics are most effective for different categories of statements and which models perform best overall.

Overall, the proposed work would involve a comprehensive analysis of different NLP metrics and their suitability for different categories of statements. It would also involve the development and optimization of NLP models for each category, and an empirical evaluation of the performance of these models using appropriate metrics.

IV. BLOCK DIAGRAM



Evaluation Process of NLP Metrics

V. ALGORITHM

Image Transformer is a state-of-the-art model for image captioning that uses self-attention and transformers to generate image descriptions. Here is an overview of the algorithm:

1. Input Encoding: The input image is encoded using a pre-trained convolutional neural network (CNN) to extract visual features from the image.

2. Text Encoding: The input text caption is tokenized and embedded using a word embedding layer to generate a sequence of word vectors.

3. Positional Encoding: Positional encoding is applied to the word embeddings to capture the relative position of the words in the sequence.

4. Self-Attention Encoder: The encoded image features and text embeddings are passed through multiple layers of self-attention encoder to generate a joint embedding. Each layer in the self-attention encoder consists of a multi-head attention layer and a position-wise feedforward layer.

5. Decoder: The joint embedding is then passed through a decoder, which generates a sequence of words one at a time. At each time step, the decoder attends to the joint embedding and generates the next word in the sequence using a softmax function.

6. Training: The model is trained using a cross-entropy loss between the predicted sequence of words and the ground truth caption.

VI. MODEL DESCRIPTION

1. CNN Encoder: A pre-trained CNN (such as ResNet or Inception) is used to encode the input image into a feature vector. The output of the CNN encoder is a 3D tensor with dimensions (N, H, W, C), where N is the batch size, H and W are the height and width of the image, and C is the number of channels (i.e., the number of filters in the last convolutional layer).

2. Word Embedding Layer: The input text caption is tokenized into a sequence of words, and each word is embedded into a high-dimensional vector space using a pre-trained word embedding layer (such as GloVe or Word2Vec).

3. Positional Encoding Layer: Since the transformer architecture does not take into account the order of the words in the sequence, a positional encoding layer is added to the word embeddings to capture their relative position.

4. Self-Attention Encoder: The self-attention encoder is a stack of N identical layers, each consisting of a multi-head attention layer and a position-wise feedforward layer. The multi-head attention layer computes a weighted sum of the input features, where the weights are determined by an attention mechanism that depends on the similarity between the input features. The position-wise feedforward layer applies a non-linear transformation to each feature independently.

5. Decoder: The decoder is another stack of N identical layers, each consisting of a masked multi-head attention layer, a multi-head attention layer that attends to the encoder output, and a position-wise feedforward layer. At each time step, the decoder attends to the encoder output and generates the next word in the sequence using a softmax function.

6. Output Layer: The output layer is a fully connected layer that maps the decoder output to the vocabulary size (i.e., the number of words in the caption vocabulary). The output of the softmax function is the probability distribution over the vocabulary, and the predicted word is the word with the highest probability.

7. Loss Function: The model is trained using a cross-entropy loss between the predicted sequence of words and the ground truth caption. During training, the model generates captions for the input images, and the loss is computed by comparing the generated captions to the ground truth captions.

VII. ANALYSIS

Descriptive captions describe the visual content of the image or video and provide relevant details about the objects, people, scenes, and actions depicted in the media. BLEU metrics figure this out better than other metrics.

Sentiment-based captions focus on the emotional content of the image or video and describe the mood, feelings, and sentiments conveyed by the visual content. SPICE metrics would be able to analyse and categorize this quantitatively.

Contextual captions consider the context of the image or video and provide information that is relevant to the specific situation or scenario depicted in the media. CIDER metrics would work better in capturing features of those textual information.

Storytelling captions use a narrative approach to describe the visual content and provide a sequence of events that tells a story about the image or video. METEOR metrics would comparatively work better at classification and accuracy analysis of these captions.

Question-answering captions: These types of captions use natural language processing and deep learning techniques to generate captions that answer specific questions about the visual content of the image or video. SPICE metric can be of use here as it captures scenic featural information better.

In linguistics, lexical similarity is a measure of the degree to which the word sets of two given languages are similar. METEOR metric qualitatively works in this case.

The syntactic similarity is based on the assumption that the similarity between the two texts is proportional to the number of identical words in them. ROGUE metric handles statements of such better.

VIII. SUMMARY

The main goal of the project is to identify suitable NLP metrics for evaluating the quality of machine-generated captions in various categories of text. This is important because different categories of text may require different NLP metrics for accurate evaluation. For example, the metric suitable for evaluating the performance of machine-generated captions for news articles may not be appropriate for social media posts or scientific papers. Therefore, it is essential to identify the appropriate NLP metrics that can effectively evaluate the quality of machine-generated captions for each category of text.

To achieve this goal, the project involves collecting diverse datasets that contain text samples from different domains and categories. The datasets are preprocessed to remove noise, normalize text, and extract relevant features. Then, a set of NLP metrics is applied to evaluate the performance of machine-generated captions against ground truth captions. The metrics used in the evaluation process include BLEU, METEOR, ROUGE, CIDEr, and others.

To ensure the reliability and validity of the experimental results, the project involves conducting multiple experiments and comparing the results of different metrics. This allows for the identification of the most suitable metrics for each category of text. The project also considers the challenges of dealing with variations in language, domain-specific terminology, and text structure.

In summary, the project aims to enhance the accuracy and quality of machine-generated captions by identifying suitable NLP metrics for each category of text. This will lead to the development of more effective and accurate NLP models that can be applied in various domains, including news articles, social media posts, scientific papers, and other forms of text.

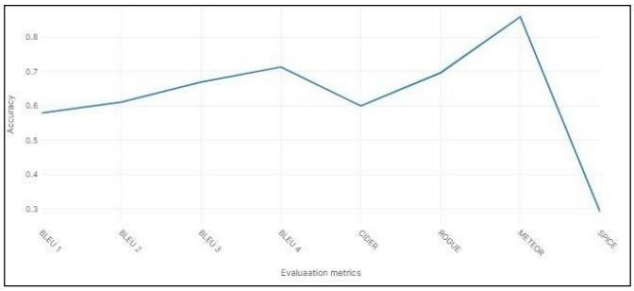
IX. RESULTS

Categories of Image Captioning and their suitable evaluation metrics

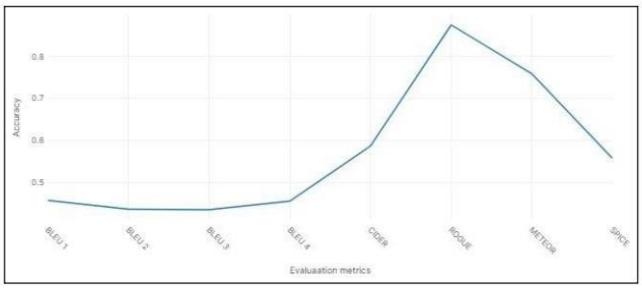
Categories of Image Captioning	Suitable Evaluation Metrics
Lexical Similar captions	Meteor
Syntactic Similar Captions	Rogue
Descriptive Captions	Bleu
Sentimental Analysis Captions	Spice
Contextual Captions	Cider
Story Telling Captions	Meteor
Question Answering Captions	Spice

Graphs for various categories of image captions and evaluation metrics

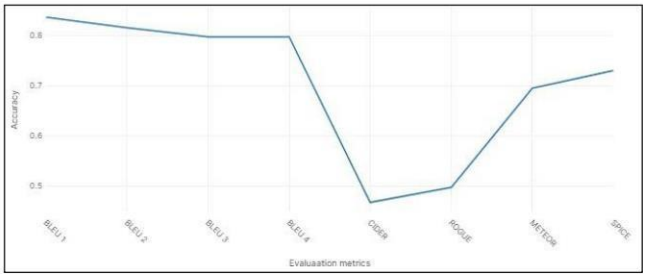
I. Lexical Similar captions



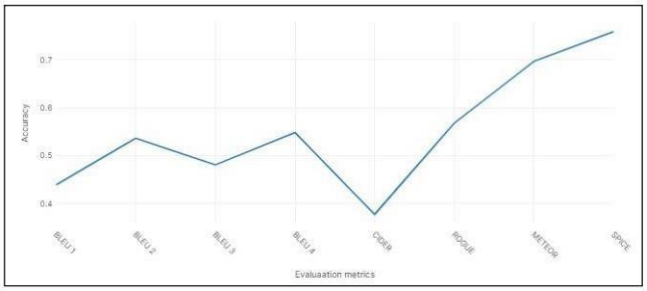
II. Syntactic Similar Captions



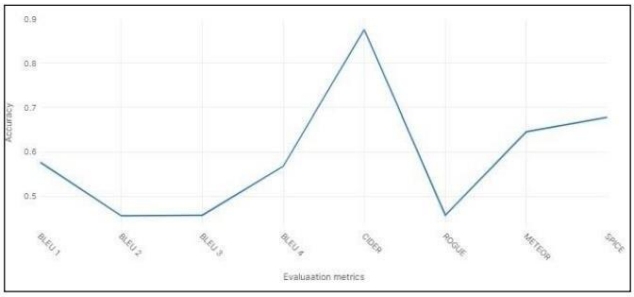
III. Descriptive Captions



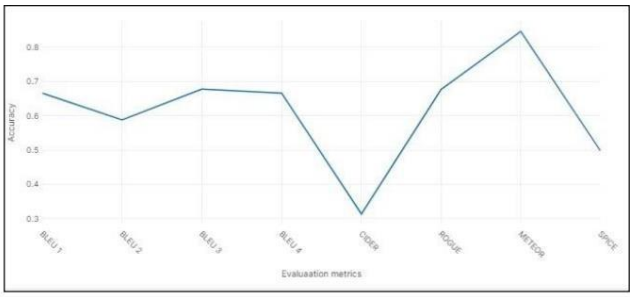
IV. Sentimental Analysis Captions



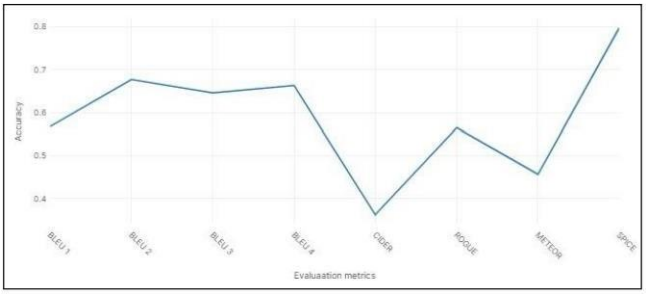
V. Contextual Captions



VI. Story Telling Captions



VII. Question Answering Captions



X. FUTURE SCOPE

The future scope of a project that aims at finding NLP metrics suitable for various categories of text is vast and promising. The project lays the groundwork for developing more sophisticated and effective NLP models that can be applied to different types of text, including social media posts, news articles, scientific papers, and other forms of text.

One potential direction for future research is to investigate the effectiveness of NLP metrics for evaluating the quality of machine-generated captions in different languages. This would involve collecting datasets in different languages and testing the performance of various NLP metrics. Another direction is to explore the use of advanced deep learning techniques such as attention-based models, transformers, and generative adversarial networks (GANs) for generating high-quality captions that are more contextually and semantically relevant.

Moreover, the project can also be extended to other areas of NLP, such as text classification, sentiment analysis, and machine translation. By identifying suitable NLP metrics for different categories of text, the project can provide a framework for developing more accurate and effective NLP models in these areas.

In summary, the future scope of this project involves further research and development in the area of NLP metrics for various categories of text, which will enable the development of more sophisticated and effective NLP models for a wide range of applications.

XI. REFERENCES

[1] 1 . Image Captioning through Image Transformer by Sen He, et al.

[2] Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional

[3] Image caption evaluation. In: European Conference on Computer Vision. pp. 382–

[4] 398. Springer (2016) 4, 9

[5] 2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.:

[6] Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern

[7] Recognition. pp. 6077–6086 (2018) 1, 4, 10

[8] 3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on

[9] Intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005) 9

[10]4. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn:

[11]Spatial and channel-wise attention in convolutional networks for image captioning.

[12]In: Proceedings of the IEEE conference on computer vision and pattern recognition.

[13]pp. 5659–5667 (2017) 1, 3

[14]5. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Doll´ar, P., Zitnick, C.L.:

[15]Microsoft coco captions: Data collection and evaluation server. arXiv preprint [16]arXiv:1504.00325 (2015) 3, 9, 13

[17]6. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated

[18] Convolutional networks. In: Proceedings of the 34th International Conference on

[19] Machine Learning-Volume 70. pp. 933–941. JMLR (2017) 5, 9, 11

[20]7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A largescale hierarchical image database. In: 2009 IEEE conference on computer vision

[21]and pattern recognition. pp. 248–255. Ieee (2009) 3, 10

[22]8. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction

[23] with lstm. Neural Computation 12(10), 2451–2471 (2000) 3

[24]On computer vision and pattern recognition.

pp. 770–778 (2016) 10