FINAL YEAR PROJECT-ZEROTH REVIEW

DECEMBER 2020 - MAY 2021

IMAGE CAPTIONING WITH SEMANTIC ATTENTION FOR VISUALLY IMPAIRED

Kavya S - 2017103543

Raveena R – 2017103577

Varsha A – 2017103606

Project Guide Signature

(Prof. Dr. V. Mary Anita Rajam)

Introduction:

Image captioning aims to provide machine-generated natural language captions for any given image. Our ability to effortlessly point out and describe all aspects of an image relies on a strong semantic understanding of a visual scene and all of its elements.

In this work, we mainly focus on improving the effectiveness of using the image attribute based semantic attention, considering that attributes contain both the high-level knowledge of the image content and specific semantics of corresponding captioning words. Image attributes are commonly chosen from a data-driven vocabulary, which is established by selecting the most commonly used words in the ground truth training captions.

The goal of this project is to help the visually impaired to understand an image with voice assistance and braille format of the generated caption. Although the visually impaired people use other senses such as hearing and touch to recognize the events and objects around them, the life quality of those people can be improved with the help of voice assisted image captioning.

Previous image attribute based methods usually leverage pre-trained deep networks to predict image attributes which are then input to the image captioning network. However, these approaches only use the classification results from attribute classification/detection, leaving the rich semantics of attributes unused.

A multimodal attribute detector (MAD) is used to effectively utilize the rich semantics of attributes. Benefiting from the rich semantic information contained in the embedding of attribute words, better image captioning results can be achieved. An ideal way of utilizing attribute words is to include only those attributes which are closely related to the next word to be generated in the context. Subsequent Attribute Predictor (SAP) predicts the most relevant attributes at each time step with the image attribute prior generated by MAD to prevent the attention module from attending to irrelevant attributes.

Overall objectives:

To generate captions with rich semantic information for any given image**,** the following steps are to be followed.

- Extraction of object features from the input image.
- Predicting the probability of image attributes to choose features with rich semantics for better captioning.
- Predicting the subsequent attributes that are highly relevant to current linguistic context for generating plausible image captions
- Conversion of generated captions to speech and braille format.

Literature Survey:

Early approaches on image captioning could be divided into two based on template matching and retrieval-based approaches. Template matching works by detecting object, action, scene and attributes in images and then fill them into a hand-designed and rigid sentence template. [1] The captions generated by these approaches are not always fluent and expressive. Retrieval based approaches first retrieve the visually similarity images from a large database, and then transfer the captions of retrieved images to fit the query image[2].

Introduction in use of Neural Networks for image classification, object detection and attribute learning motivated the use of Neural Networks for image captioning. The neural-network-based approaches for automatic image captioning fall into two categories.

The first one is the encoder-decoder based framework adopted from neural machine translation where the encoder RNN reads the source sentence and transforms it into a rich fixed-length vector representation which in turn is used as the initial hidden state of a "decoder" RNN that generates the target sentence[3].

However, traditional RNNs deal with the sequence in a recurrent way, squeezing the information of all previous words into hidden cells and updating the context information by fusing the hidden states with the current word information. This may miss the rich knowledge too far in the past.[4]

In order to boost the performance of encoder-decoder, visual attention was introduced. Visual attention utilizes the extracted spatial feature or object feature to describe the image. The visual attention mechanism was later introduced to improve the accuracy of the generated captions by effectively utilizing the visual features.[5]

The other category of work is based on a compositional approach which employs a CNN to detect a set of semantic tags, then uses a maximum entropy language model to generate a set of caption candidates, and finally adopts a deep multimodal similarity model to re-rank the candidates to generate the final caption. [6]

Replacing the encoder RNN with Deep Convolutional Neural Network (CNN) to produce a rich representation of the input image by embedding it to a fixed-length vector, such that this representation can be used for a variety of vision tasks thereby improving the performance.[7]

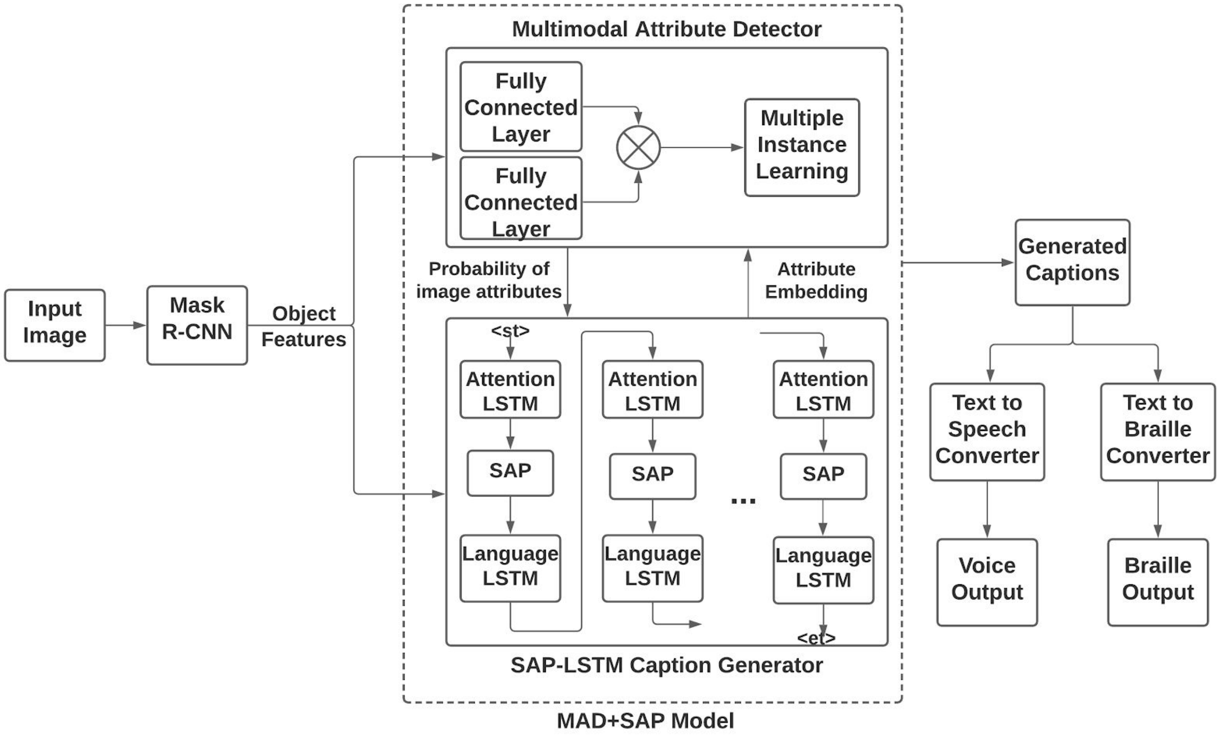
To address the localization of images and description [8] proposes a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization.

Z. Gan et al [9] proposed a Semantic Compositional Network (SCN) is for image captioning, in which semantic concepts are detected from the image, and the probability of each semantic concept is used to compose the parameters in a long short-term memory (LSTM) network. The SCN extends LSTM for ensembling the semantic concepts .

[10] Proposes a comprehensive text to speech synthesis technology combining Natural Language Processing(NLP) and Digital Signal Processing (DSP). NLP does pre-processing, morphological analysis, contextual analysis, syntactic-prosodic analysis, phonetization and prosody generation. DSP does two types of synthesis methods which are rule-driven methods and data-driven methods. Text to speech synthesis module is used to convert the generated caption into audio format.

[11] presented a prototype for portable image to braille display conversion using solenoids and servo motors as actuators and microcontroller to convert each character of text into braille script. In order to assist visually impaired to understand what the image depicts, this project converts the generated caption into braille format which can be then printed for future references.

Block diagram:



The object features are extracted from the input image using Mask R-CNN and is fed to both Multimodal Attribute Detector (MAD) and Subsequent Attribute Predictor - Long Short Term Memory (SAP-LSTM) caption generator.

MAD uses fully connected layers and matrix multiplication along with Multiple Instance Layer (MIL) to compute the probability of whether an attribute is contained in the input image.

The Attention LSTM helps in prediction of subsequent attributes and distinguishes the importance of different object features.

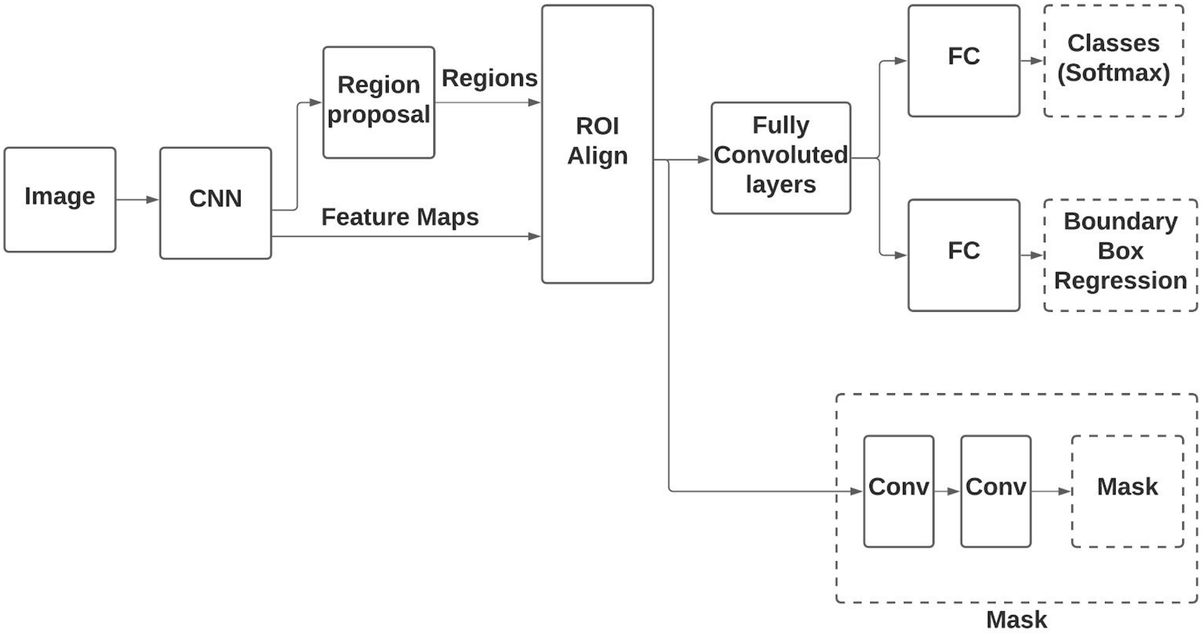
SAP uses the probability computed by MAD and the output of Attention LSTM to choose the most relevant subsequent attribute at each time.

The subsequent attributes of top-k probabilities chosen in SAP are fed to Language LSTM to check the semantics of the generated partial caption and a semantically accurate caption is generated.

The caption generated with semantic attention is given to text to speech converter and text to braille converter which converts the caption to voice and braille format so that the visually impaired will be able to understand any image.

Details of modules:

Mask R-CNN:

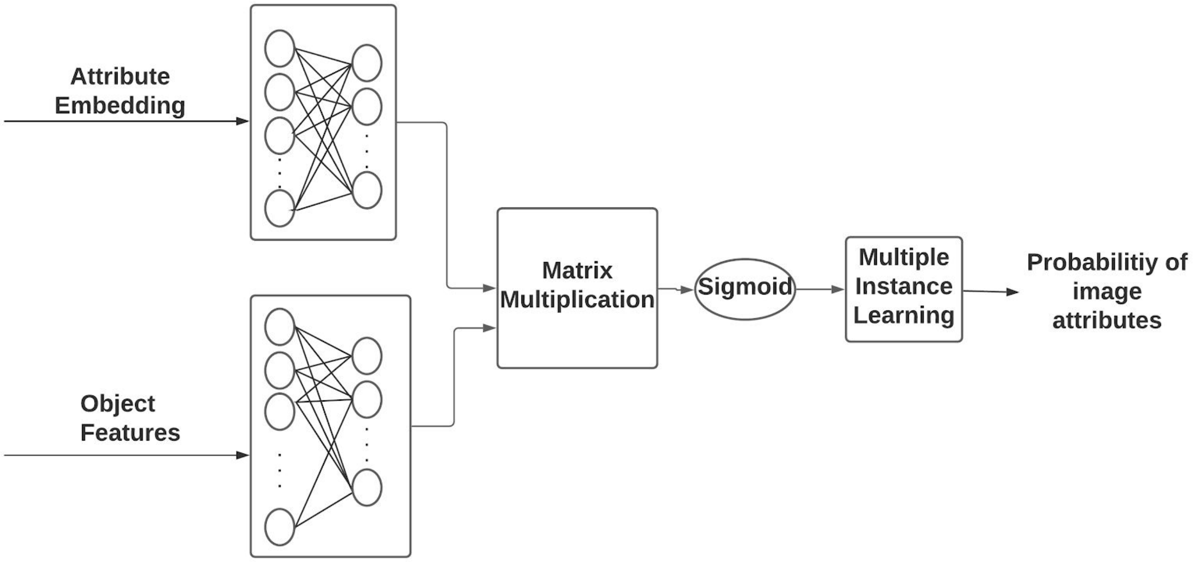


Mask R-CNN is a state of the art model for instance segmentation, developed on top of Faster R-CNN. Faster R-CNN is a region-based convolutional neural network that returns bounding boxes for each object and its class label with a confidence score. It predicts object class and bounding boxes.

Mask R-CNN is an extension of Faster R-CNN with additional branches for predicting segmentation masks on each Region of Interest (RoI). RoIAlign helps to preserve spatial information and uses binary interpolation to create a feature map that is of fixed size.

The output from the RoIAlign layer is then fed into Mask head, which consists of two convolution layers. It generates a mask for each RoI, thus segmenting an image in pixel-to-pixel manner.

Multimodal Attribute Detector:

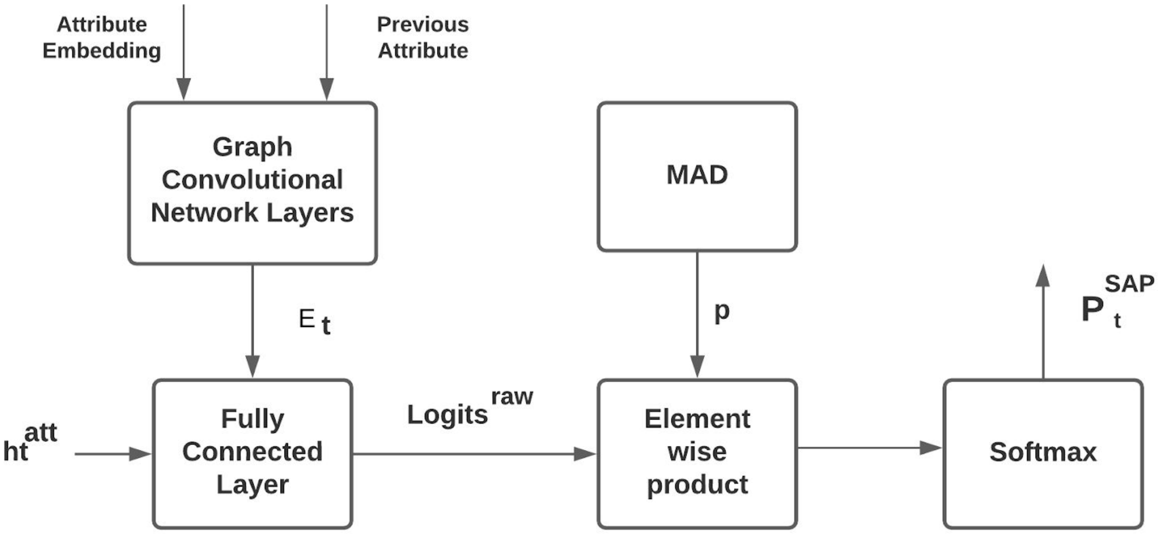


The **Multimodal Attribute Detector** predicts the probability of whether an attribute is contained in the input image using two steps:

1. The attribute embedding and object features are mapped to the same space using two fully connected layers. The outputs are merged using matrix multiplication to compute the similarity. The product is sent to the sigmoid layer which gives a raw probability matrix P_{raw} where P_{raw}^{ij} denotes the probability that j^{th} object contains the ith attribute a_i .

2. The raw probability values in each row of P_{raw} are merged using noisy OR Multiple Instance Learning to predict the final probability p_i that the input image contains the i^{th} attribute a_i.

Subsequent Attribute Predictor:



The **Subsequent Attribute Predictor (SAP)** predicts an appropriate subset of attributes, or subsequent attributes, to attend at each time step. It keeps the most relevant attribute at each time step and discards the other attributes.

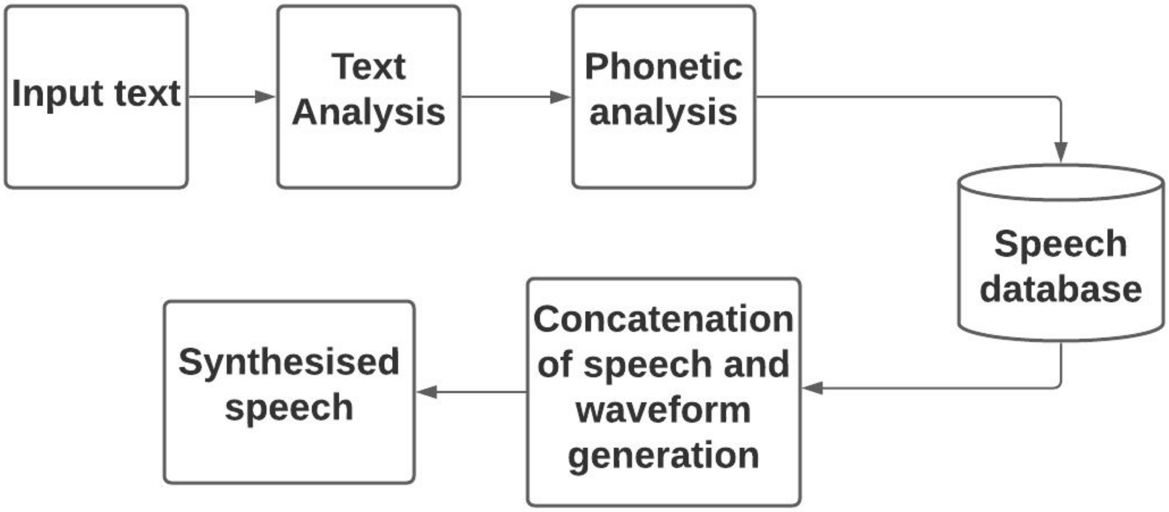
The **Graph Convolutional Network layer** updates the embedding of attributes in order to refine the features of the attributes. It takes the previous attribute and transition probability matrix and updates the embedding attribute using Leaky Relu activation function. The transition probability matrix is obtained by pairing the neighbouring attributes of the ground truth training captions.

The **Fully Connected Layer** uses the output of Attention LSTM to distinguish the importance of different object features and the output of the GCN layer to generate $logits^{raw}$ which is an intermediate output.

The **Element wise product** layer uses probability p from MAD as weights along with $logits^{raw}$ to generate probability distribution of subsequent attributes.

Instead of choosing one attribute with maximum probability, top-K attributes are chosen as current subsequent attributes in order to prevent the model from using the wrong attribute when the top-1 subsequent attribute is erroneous.

Text to speech converter:

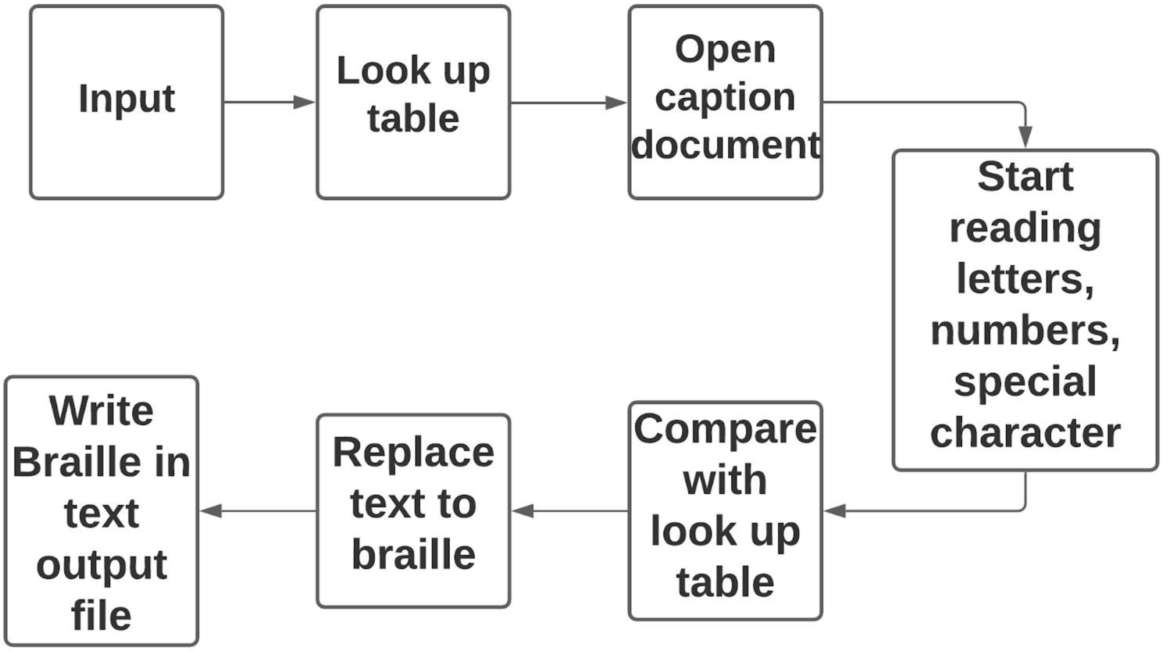


The Text-to-Speech synthesis procedure consists of two main phases.

1. The first one is text analysis, where the input document structure is detected, text is normalized and linguistic analysis is performed. Phonetic analysis identifies the homograph disambiguation and converts the grapheme to phoneme.

2. The second one is the generation of speech waveforms. The speech signal is generated by concatenating pre recorded sound units from the speech database and synthesised speech is generated as output.

Text to braille converter:


Input
Look up table
Open caption document
Start reading letters, numbers, special character
Compare with look up table
Replace text to braille
Write Braille in text output file

The input text indicates a word (.doc) or text (.txt) document that contains letters in the English language along with numbers and special Characters.

In the Lookup table (LUT) all the braille codes for the 26 letters, 10 numbers and various special characters are written in an array or a table format.

Compare the letters with the LUT and convert them accordingly into Braille by switching text.

After the completion of the conversion process, the braille contents are written in the output file.

Performance measures:

Performance is measured in terms of:

1. **BLEU-4**

   BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. “The closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU.

   In this, the geometric mean of the test corpus’ precision score is taken and then multiply the result by an exponential brevity penalty (BP) factor. We first compute the geometric average of the modified n-gram precisions, p_n, using n-grams upto length N and positive weights w_n summing to one. Let c be the length of the candidate translation and r be the effective reference corpus length. To compute the brevity penalty BP,

$$BP = \begin{cases} 1 & \text{if } c>r \\ e^{(1-\frac{r}{c})} & \text{if } c\leq r \end{cases}$$

$$BLUE = BP.exp\left(\sum_{n=1}^{N} w_n log\, p_n\right)$$

2. **METEOR**

   METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric for the evaluation of machine translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.

   To account for gaps and differences in word order, a fragmentation penalty is calculated using the total number of matched words and number of chunks (ch).

$$P_{en} = \gamma.\left(\frac{ch}{m}\right)^{\beta}$$

The Meteor score is calculated by

$$Score \;=\; (1 - P_{en})\,.F_{mean}$$

The parameters β, γ are tuned to maximize correlation with human judgments.

3. ROUGE

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics used for evaluating automatic summarization and machine translation software in NLP. The metrics compare an automatically produced summary or translation against a reference (human-produced) summary or translation.

ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as

$$ROUGE - N = \frac{\sum\limits_{S \in \{ReferenceSummaries\}} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in \{ReferenceSummaries\}} \sum\limits_{gram_n \in S} Count(gram_n)}$$

where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries.

4. CIDER-D

CIDEr is Consensus-based Image Description Evaluation. A version of CIDEr named CIDEr-D is available as a part of MS COCO evaluation server to enable systematic evaluation and benchmarking.

$CIDEr_n$, score for n-grams of length n is computed using the average cosine similarity between the candidate sentence and the reference sentences, which accounts for both precision and recall.

$$CIDEr_n(c_i,\ S_i) \ = \ \frac{1}{m}\sum_{j} \frac{g^n(c_i)\ .\ g^n(s_{ij})}{||g^n(c_i)||\ ||\ g^n(s_{ij})||}$$

5. **SPICE** (Semantic Propositional Image Caption Evaluation)

SPICE captures human judgments over model-generated captions better than other automatic metrics (evaluated over a range of models and datasets)

SPICE measures how well caption generators recover objects, attributes and the relations between them.

The Precision P, recall R, and SPICE are defined as

$$P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$$

$$R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$$

$$SPICE(c, S) = F_1(c, S) = \frac{2\ .\ P(c,S)\ .\ R(c,S)}{P(c,S) + R(c,S)}$$

Dataset:

This project evaluates model performance on the MS-COCO captioning dataset. MS-COCO is a large-scale object detection, segmentation, and captioning dataset. It uses 5000 images for validation and 5000 images for testing from the 40504 validation set following the widely adopted Karpathy's data split.

References:

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. "Every picture tells a story: Generating sentences from images." In ECCV, pages 15–29, 2010.

[2] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47:853– 899, 2013

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3156–3164

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016. pp. 4651 - 4659

[5] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 3, no. 5, Jun. 2018, p. 6.

[6] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From ´ captions to visual concepts and back. In CVPR, pages 1473– 1482, 2015

[7] Hartati,Hanif Al Fatta, and Utsman Fajar, “Captioning Image Using Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM)” in IEEE International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2019, pp. 263-268

[8] Justin Johnson Andrej Karpathy Li Fei-Fei, “DenseCap: Fully Convolutional Localization Networks for Dense Captioning” in Proc. Dec 2016 IEEE Conference on Computer Vision and Pattern Recognition.pp. 4565-4574

[9] Z. Gan et al., “Semantic compositional networks for visual captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 2, Jul. 2017. pp. 1141 - 1150

[10] Zhigang Yin, “An Overview of Speech Synthesis Technology”, in Proc. 2018 International Conference on Instrumentation, Measurement, Computer, Communication and Control (IMCCC), Mar. 2020, pp. 522-526

[11] Kimaya Kulkarni Apoorva Mahajan Yash Zambre “Text Detection and Communicator Using Braille for Assistance to Visually Impaired” in Conf.,2019 IEEE Pune Section International Conference (PuneCon), pp. 1-5