Adama Science and Technology University

School of Electrical Engineering and Computing

Department of Computer Science and Engineering

Research Methodology and Ethics(CSE-6024)

Masters of Science in Computer Science and Engineering(MSc.)

Research Proposal

Real-time Voice Synthesis from an Image for Physical World Recognition by using Machine Learning

By

Belay Birhanu G. PGR/24497/14

Melaku Yilma A. PGR/24625/14

Mihretu W/Yohannes PGR/24515/14

Milkesa Abdi K. PGR/24510/14

Advisor: Dr. Teklu Urgesa

May 2022,
Adama, Oromia, Ethiopia

Abstract

We have proposed a model that blends various machine learning categories and converts visual words to auditory in this proposal. The key issue we wanted to address was the difficulty that visually impaired people had in their daily lives. To generate voice from an image in real time, we will combine a state-of-the-art CNN model with two RNN models. The first portion of our model, CNN, extracts features from an image, while the remaining two RNN models, LSTM with attention and Seq2Seq, generate natural language (in English) description for images and translate them to Afaan Oromo. Finally, we turn the natural language description of an image into audio using Google's voice synthesis model.

Key Words: CNN, RNN, LSTM, Seq2Seq

Table of Contents

Abstract i
1. Introduction 1
2. Background 1
3. Problem Statement 2
4. Research questions 2
5. Objectives 2
5.1 Main objectives 2
5.2 Specific objectives 2
6. Definition of key terms 3
7. Literature Review 3
8. Methodology 6
8.1 Study design 6
8.2 Study Setting 8
8.3 Plan for data analysis 8
8.3.1 Datasets 8
8.3.2 Study variable 9
8.3.3 Data collection tools and procedures 9
8.3.4 Simulations 9
9. Budget 9
10. Timelines 10
11. References 11

1. Introduction

Many factors influence a person's perception of vision impairment. This can include things like the unavailability of prevention and treatment strategies, vision rehabilitation (including assistive devices like spectacles or white canes), and whether the person has trouble using inaccessible buildings, transportation, or information. According to recent records of WHO, near- or far-sightedness affects at least 2.2 billion individuals worldwide. Vision impairment might have been avoided or managed in at least 1 billion of these cases (almost half).

Early-onset severe vision impairment can cause delays in motor, verbal, emotional, social, and cognitive development in young children, with long-term repercussions. School-aged children with vision impairment may struggle academically.

People need visual information to interpret the world. The scene images can be converted into their own cognition. People can gradually understand the world around them by accumulating information. Image captioning is a research area that turns images into knowledge. Its fundamental model requires two functions. The first phase involves extracting image feature information, which mostly entails extracting object information and object position information from a picture; the second half involves analyzing image semantic information and combining it with image features to provide an image description. The human brain is equipped with a full cognitive system.

The brain will process and interpret the image content as long as the image is received. When a computer realizes image captioning, it is frequently required to infuse cognitive image capability into the machine. This function cannot be achieved with a regular program. On one hand, the logic to be considered is too complicated, and the program is too huge; on the other hand, the typical program is too inflexible to produce the desired effect. The neural network algorithm is used to get the computer closer to people's cognitive abilities and to achieve the language description level of children.

In recent years, the idea of automatically creating descriptive words for images has sparked attention in natural language processing and computer vision research. Image captioning is a critical endeavor that requires both a semantic understanding of images and the ability to create correct and precise description phrases.

2. Background

Image captioning is one of the hottest topics in artificial intelligence. It has a wide range of applications, including human-computer interaction, adding subtitles to videos, answering video questions, searching for important information based on image content, and image search by keywords, among others. The application of image captioning in road condition detection allows the visually impaired to see the external environment in real time, providing tremendous convenience and safety for travel.

Templates are used to implement the original image captioning algorithm. This method collects a sequence of important feature data, such as key objects and special attributes, using various types of classifiers, such as SVM, and then converts the feature data into description text using lexical models or other particular templates. The object in the image, the object's action, and the scene in which the object is positioned are the three basic features extracted. Some typical smoothing

methods remove the noise component. Finally, the link between the retrieved data and the final picture description result is examined.

In terms of prediction, the image captioning model based on the encoder decoder structure is more flexible. Convolution neural networks are used in the encoder to extract and vectorize the major image feature information; recurrent neural networks are used in the decoder to combine the vectorized image features with semantic information to construct an image content description statement. According to recent researches, combined information from spatial and channel level attention is used to construct visual description. This model differs from the previous attention method in that it can use the collected text sequence's context semantic information at the encoding end.

3. Problem Statement

A person's eyes are their primary sensory organ. A quick glance around us reveals how visual most of the information in our environment is. Timetables in train stations, signs indicating the right way or potential danger, a billboard advertising a new product on the market are all examples of visual information that we encounter on a daily basis. The majority of this information is inaccessible to the blind and visually impaired, limiting their independence because access to information represents autonomy. Adults' quality of life is significantly impacted by vision impairment. Adults with low vision often have lower rates of workforce participation and productivity, as well as higher rates of depression and anxiety. In the case of older adults, vision impairment can lead to social isolation, difficulty walking, a higher risk of falls and fractures, and other problems.

4. Research questions

How can we translate visual information to audible information?
What sort of machine learning application can we develop to make the visualize world audible?
How can we help visually challenged individuals understand everything around them using cell phones?
How can we help a strange who has limited knowledge of the Afaan Oromo understand the local names of objects without any guide?

5. Objectives

5.1 Main objectives

The major goal of this study is to use our understanding of computer vision and natural language processing to develop a machine learning application that will synthesize voice from an image, allowing the visual world to be heard.

5.2 Specific objectives

We are going to develop an application that is capable of:

- ✓ Extracting image features using CNN model
- ✓ Generate natural language (in English) description of an image by using LSTM with attention model.
- ✓ Translate the generated description or caption to Afaan Oromo using Seq2Seq model
- ✓ Synthesis voice from the translated text (Afaan Oromo)

6. Definition of key terms

CNN: Convolutional Neural Networks are powerful image processing, artificial intelligence (AI) that use deep learning to perform both generative and descriptive tasks, often using machine vison that includes image and video recognition, along with recommender systems and natural language processing (NLP).

RNN: Recurrent Neural Networks are a class of neural networks that are helpful in modeling sequence data. Derived from feedforward networks, RNNs exhibit similar behavior to how human brains function. Simply put: recurrent neural networks produce predictive results in sequential data that other algorithms cannot.

LSTM: Long short-term memory is an artificial neural network used in the fields of artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Such a recurrent neural network can process not only single data points (such as images) but also entire sequences of data (such as speech or video).
A common LSTM unit is composed of cell, an input gate, an output gate and forget gate. The cell remembers values over arbitrary time interval and the tree gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing, and making predictions based on time series data, since there can be lags of unknown duration between important events in time series.

RNN Encoder-Decoder: RNN Encoder-Decoder, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair. The encoder maps a variable-length source sequence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length target sequence.

seq2seq model: Sequence to Sequence models is a special class of Recurrent Neural Network architectures that we typically use (but not restricted) to solve complex Language problems like Machine Translation, Question Answering, creating Chatbots, Text Summarization, etc.

7. Literature Review

Image caption generation has been investigated for a long time as a way to connect computer vision with natural language processing. Retrieval based system in which a user provide image as input and retrieval-based model retrieve a candidate caption that best describes the query image from a pre-constructed repository were used as an early image captioning method. The Models based on retrieval can generate captions that are both informative and grammatically correct. Despite this, these models frequently struggle to come up with creative and different captions that accurately reflect the new images. Although retrieval-based methods can produce syntactically correct captions, they are not tailored to the query images and are limited by the size of the pre-constructed image-caption repository [4] [5].

The seq2seq model, which uses CNN as the encoder and (LSTM) as the decoder, is a standard generation-based image captioning method. The encoder will use one or a few distributed and real-valued vectors to capture the semantics and content of the query image; the decoder will convert the query image's distributed vectors to a textual image description. Because they can be taught end-to-end and scale to large-scale training data, seq2seq-based techniques have become the industry standard in image captioning. Furthermore, these algorithms can generate a fresh sentence as the caption and produce outcomes that are both flexible and high-quality. However, it is

generally known that generation-based models are prone to producing generic, uninformative, and grammatically wrong captions [3] [4]. The capacity of the present image-caption repository limits the number of captions that may be accessed. Because the retrieved caption is not personalized for the query image, even the best-matched caption from the repository is not guaranteed to be a decent caption [4] [8]. Later on, researchers, [2], proposed two methods: visual attention and semantic attention to describe the image comprehensively by making use of the extracted spatial feature and focus on image details by using image attributes.

Moreover, deep learning has recently been widely used in image caption generation tasks, yielding excellent results. According to [7] and [8], the "encoder-decoder" process was used, in which the encoding stage used convolutional neural networks (CNNs) to extract visual features of the image, and the decoding stage took the extracted image visual features as input and used recurrent neural networks (RNNs) to generate the words of the image description one by one.

[1] proposed Attentive Linear Transformation model to extract significant information from the image feature space to the context vector space. The caption model can use the proposed ALT to combine spatial attention, channel-wise attention, visual dependence, and other important high-level semantics image captioning. For computing attention probabilities over image regions, they present a soft threshold regression, which outperforms softmax regression.

With the goal of improving image feature extraction and the attention mechanism, [7] proposed an image captioning model based on DenseNet and the adaptive attention mechanism. To extract more detailed global image features, the DenseNet network is used. The adaptive attention mechanism can then decide when to rely solely on visual signals and when to rely solely on the language model.

Long-term recurrent convolutional network for visual description and multimodal recurrent neural network (MRNN) for sentences generation has bene been proposed by [6] for image captioning, however these models have two major flaws. First, the majority of them are built using a single neural network, which is inflexible when it comes to learning deep semantics representation. Second, they solely extract unidirectional syntactic aspects from texts, disregarding bidirectional context information [5]. Two overcome these problems, [5] proposed cascaded recurrent neural network (CRNN). Unlike typical MRNN models, CRNN has a front-end and a back-end network that are linked to extract visual-language interactions from bi-directions. In this proposed model, a stacked Gated Recurrent Unit (SGRU) is made up of two hidden layers that expand the recurrent neural network's vertical depth, allowing it to find semantic connection between images and phrases. CRNN consists of three parts: CNN for image feature extraction, a front-end network for extracting image-word mapping and a back-end network for exploiting deep semantic context.

On the translation side, the NMT model, which is effectively a single huge neural network in concept, is built entirely by learning translation expertise from the training corpus with minimal effort from linguists or engineers. Second, in the machine translation process, gating and attention techniques, which are widely used in the field of Natural Language Processing (NLP), have been shown to be effective in capturing potential long-distance dependencies and complicated word alignment information, both of which are serious challenges for SMT. Finally, the NMT model uses substantially less memory than traditional SMT models, which keep a huge translation model, a reordering model, and a language model in memory [10] [11] [12] [13].

Despite these benefits, recent research suggests that NMT models provide fluent but occasionally erroneous translation sequences due to the following limitations: NMT completes target sentence production by creating the 'EOS' end-of-sentence signal. NMT has a coverage problem in the sentence generation process, where it cannot guarantee that all source words are accurately translated, resulting in either "overtranslation" or "undertranslation" [12]. The NMT decoder is simply a language model that is conditioned on the source sentence representation. As a result, NMT usually yields words that are considered to be very fluent, but they may not convey the encoded source sentence's original meanings [14] [15]. Because target word prediction in the softmax layer is computationally expensive, NMT uses a limited vocabulary to record the most frequent terms and a UNK symbol to represent other words on both language sides. The presence of UNK symbols makes it harder to translate these uncommon terms [14].

In comparison to NMT, SMT models use a different translation process that can effectively address the aforementioned flaws: when all source segments are covered by the translation rules, the SMT decoder completes the translation process. It ensures that every word in the source sentence is translated; SMT models translate a source word to a semantically related target word, reducing the problem of imprecise translation; In SMT translation rules, words are explicitly learned. As a result, the SMT translation rules do not include any UNK symbols [10].

To cope with these problems, [10] proposed a model that integrate statistical machine translation and neural machine translation. To incorporate SMT word knowledge into the NMT neural network-based architecture, an additional neural network-based classifier is used to score and estimate the probabilities of word recommendations from the SMT model, and then these word recommendation probabilities are used to adjust the NMT word prediction probabilities [10].

LSTM with attention model has shown promising accuracy in generating semantically proper caption for an image, according to all of the articles above. The combined model of statistical machine translation word knowledge and neural machine translation model, on the other hand, has produced remarkable overall performance in language translation. None of the models above, however, coupled the caption generating and language translation models. Based on this, we proposed combining three models: caption generation using LSTM with attention, language translation using encoder-decoder with attention, and voice generation using Google's voice synthesis model. The proposed model is shown in Figure 3 below.

8. Methodology

8.1 Study design

Image captioning is describing an image using one or more natural language sentences. The requirement to translate between two separate, but frequently paired, modalities is at the heart of this topic, which combines computer vision with natural language processing. Our main goal, however, is not limited to this. Rather, we will add two more natural language processing elements to this image captioning method: neural language translation (English to Afaan Oromo) and text to voice conversion, making the visual world audible.

The goal of this research is to create a machine learning application that will take an image as input, generate a natural language description of the image, translate the description to Afaan Oromo, and generate voice from the translated language, allowing visually impaired people to recognize the physical world using their cell phone. We will combine our expertise in computer vision and natural language processing.

We can model our experiment mathematically as:

$$I = f(x, y)$$

$$f_{Cl} = CNN(I, [-2])$$

$$IC\_Map = MAP\left(f_{Cl}, C_{training}\right)$$

$$C_{caption} = LSTM_{attention}(IC\_Map)$$

$$T_{translated} = EncoderDecoder\left(C_{caption}\right)$$

$$y_{voice} = G_{generator}(T_{translated})$$

Where

- ✓ $I \rightarrow$ the input image (visual data)
- ✓ $\boldsymbol{f_{Cl}}$ is fully connected layer of the CNN model
- ✓ $\boldsymbol{CNN}(.) \rightarrow$ one of the state-of-the art modes of CNN, ex. VGG16
- ✓ $\boldsymbol{C_{training}}$ and $\boldsymbol{C_{caption}}$ are the training and the output captions (***in English***) respectively
- ✓ $\boldsymbol{LSTM_{attention}}(.) \rightarrow$ long short term memory model
- ✓ $\boldsymbol{T_{translated}} \rightarrow$ translated caption (***in Afaan Oromo***)
- ✓ $\boldsymbol{EncoderDecoder}(.) \rightarrow$ sequence to sequence translator model
- ✓ $\boldsymbol{G_{generator}}(.) \rightarrow$ voice generator model
- ✓ $\boldsymbol{y_{voice}} \rightarrow$ the final audible information

In our proposed model, as we can see from the equation above, the first step that we go through is to extract the image feature. The feature that we are interested in is the ***fully connected*** layer of CNN. Because, the last layer of CNN is the softmax or prediction layer and the layer of our interest

is a large vector that is formed by concatenating all the elements of the MaxPooling layer before it. The next step is to map these image features with the corresponding preprocessed captions. There might be more than one caption for an image, therefore, we have one-to-one or more $(1-\text{to}-1^{\cdots *})$ mapping i.e. list of captions are mapped to a single image. Then we will feed this image feature-caption mapped data to the LSTM with Attention model.

The LSTM with attention model outputs description of an image in English language. Then we will input this description into sequence to sequence (Encoder-Decoder) model to generate the translated (Afaan Oromo) version of the description. In order to maintain the long term dependencies between words or sentences we will apply the attention mechanism to the Encoder-Decoder model. From the fig.2 below, C_i is context vector that stores the long term dependencies between words.

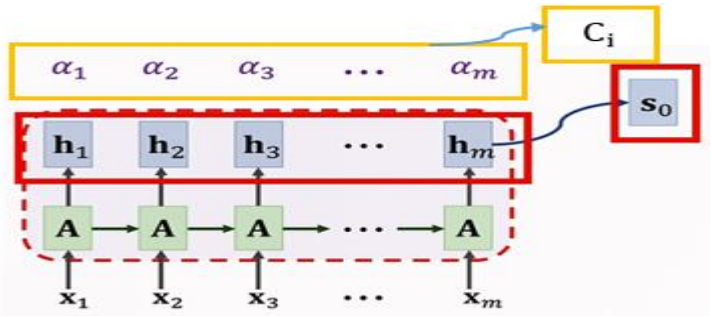


Figure 1: The Encoder model

Mathematically,

$$\hat{a}_i = V^T \cdot \tanh\left(W \cdot \begin{bmatrix} h_i \\ s_i \end{bmatrix}\right)$$

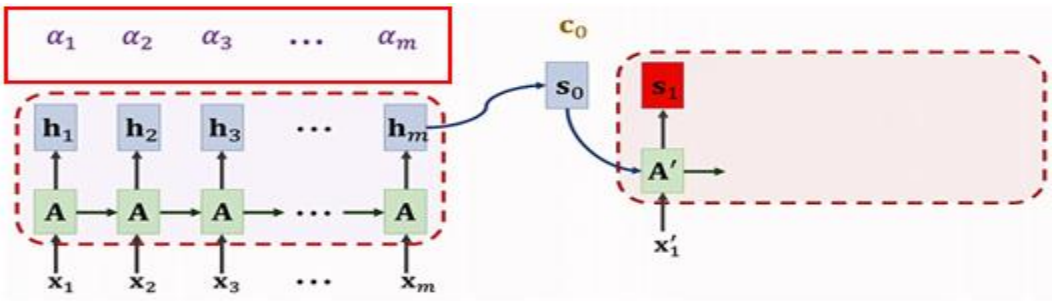
$$a_i = \text{softmax}(\hat{a}_i)$$

$$C_i = \sum_i a_i h_i,$$

Where

- ✓ $h_i \rightarrow$ hidden states
- ✓ $s_i \rightarrow$ previous output word
- ✓ V and W are trainable parameters
- ✓ $\tanh(.)$ and $\text{softmax}(.)$ are nonlinear activation functions
- ✓ $C_i \rightarrow$Context vector that the information about the previous words

Our translator model will work as follows:



As it can be seen from the figure output we feed English language to the encoder, the encoder then figures out the feature of the words and condense it into one and send to the decoder, the decoder then uses this condensed information as input along with context vector and outputs our desired translated word. i.e. x_i is English word input to the encoder, s_0 is the condensed representation of the whole input, x_i' previously generated Afaan Oromo word and it is also input to the decoder and s_i the final output (in Afaan Oromo).

S_i is computed as: $S_i = \tanh\left(A' \bullet \begin{bmatrix} x_i' \\ s_i \\ c_i \end{bmatrix} + b\right)$, where b is a bias.

The overall proposed model is:

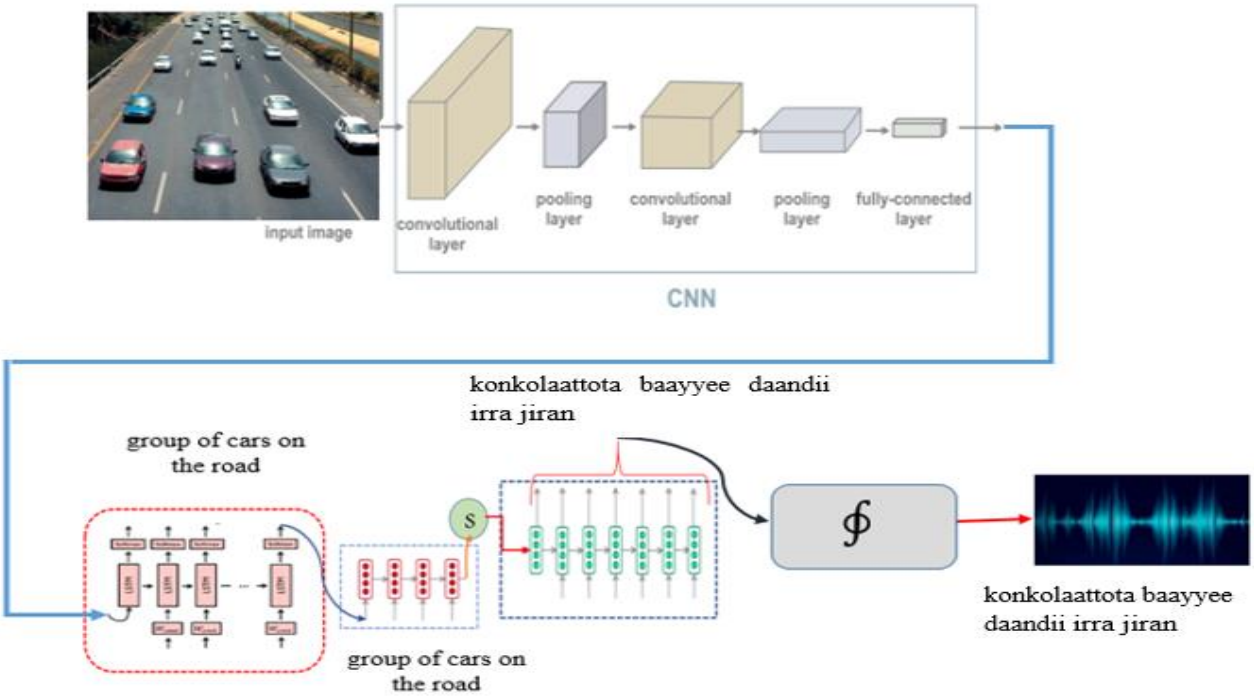


Figure 2: The Proposed Model

8.2 Study Setting

We will construct the application in a computer laboratory on internet connected computer, as indicated above, because the goal of this research is to develop a machine learning application that is capable of synthesizing voice from an image in reat-time.

8.3 Plan for data analysis

8.3.1 Datasets

Artificial intelligence is built on the foundation of data. Many difficult-to-find rules are increasingly being discovered by analyzing vast amounts of data. There are currently rich and colorful datasets in the image description generating work, such as MSCOCO, Flickr8k, Flickr30k, PASCAL 1K, AI Challenger Dataset, and STAIR Captions, and they are increasingly becoming a trend of debate. The dataset employs alternative syntax to describe the same image in order to have

many independent descriptions. Different descriptions of the same image emphasize different features of the scene or employ various grammars.

8.3.2 Study variable

Our model combines two deep neural networks - Convolutional Neural Network and Recurrent Neural Networks (LSTM with Attention). In both cases, number of metrics have to be considered while training so as to improve the overall performance of the model. Our model's most important variables are: **image-caption mapping (X)**, label (y), **loss:** measure of the deviation between the expected value and the actual value, **accuracy:** identify relationships and patterns between variables in a dataset based on the input, or training, data, and validation accuracy.

8.3.3 Data collection tools and procedures

On massive amounts of data, neural networks created the foundation for their success. As a result, we must train our model with a significant amount of data in order to get virtually exact results. To maintain adequate accuracy, we need over 10K images on average. 80 percent of the data is utilized for training, 10% is used for testing, and 10% is used for validating the model. However, we are compelled to use secondary data in certain instances, such as a shortage of data collection material, time, and so on. Furthermore, because secondary data is more structured and cleaned than primary data, it will provide us with greater accuracy.

8.3.4 Simulations

The initial draft of our experiment's simulation procedure will be completed on an internet-connected computer with a minimum of 4GB RAM. We test our model by feeding it with new images once we have fitted it to the dataset we are going to utilize. After we get a guarantee from our model, we integrate it with ReactJS and test it in real time on our phone.

9. Budget

No	Item	Description	Amount Requested	Estimated Unit Price
1.	Hard disk	1TB external hard disk	01	2500.0 ETB
2.	Travel	To collect data	-	2000 ETB
3.	Materials and supplies	Stationary materials: pen, pencil, paper and etc.	01 dozen each	200.0 ETB x 2 + 500.0 ETB
4.	Consultant fee	Consultation	1 consultant	3000.0 ETB
5.	Printing	Questionnaire papers	15	1.5 ETB
6.	Total			8401.5 ETB

10. Timelines

No.	Tasks	April 1st -5th	April 6th - 21	April 23 -13 May 2022	May 14-20	May 21-30	June 1st -5th	June 6th – 10th
1.	Data collection							
2.	Data pre-processing							
3.	Model designing							
4.	Model training							
5.	Model testing							
6.	Model Integration							
7.	Model deployment							

11. References

[1] S. Ye, N. Liu and J. Han, "Attentive Linear Transformation for Image Captioning," in Proceedings of Journal of L ATEX Class files, Vol. 14, No. 8, August 2015

[2] Y. Huang, J. Chen, W. Ouyang, W. Wan and Y. Xue, "Image Captioning with End-to-End Attribute Detection and Subsequent Attributes Prediction," on IEEE transactions on Image processing, Vol. 29, 2020.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan ,"Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge," on IEEE transactions on pattern analysis and machine intelligence, 2016.

[4] M. Yang, J. Liu, Y. Shen, Z. Zhao, X. Chen, Q. Wu and C. Li, "An Ensemble of Generation- and Retrieval-Based Image Captioning with Dual Generator Generative Adversarial Network," on IEEE transactions on Image processing, Vol. 29, 2020.

[5] J. Wu and H. Hu, "Cascade recurrent neural network for image caption generation," in the Proceedings of Electronics letters 7th December 2017, Vol. 53, No. 25, pp. 1642–1643.

[6] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell," Long-term Recurrent Convolutional Networks for Visual Recognition and Description," on IEEE transactions on pattern analysis and machine intelligence, 2016

[7] "Image captioning using DenseNet network and adaptive attention,"

[8] Z. Deng, Z. Jiang, R. Lan, W. Huang and X. Luo, "Image Caption Generation via Unified Retrieval and Generation-Based Method," in Signal Processing: Image Communication, 15 March 2020.

[9] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," September 2014.

[10] X. Wang, Z. Tu, and M. Zhang, "Incorporating Statistical Machine Translation Word Knowledge into Neural Machine Translation", Journal of LATEX Class files, Vol. 14, No. 8, August 2015.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol.9, no. 8, pp. 1735–1780, 1997.

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proceedings of ICLR 2015, 2015.

[13] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage for neural machine translation," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2016, pp. 76–85.

[14] P. Arthur, G. Neubig, and S. Nakamura, "Incorporating discrete translation lexicons into neural machine translation," in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2016.

[15] Z. Tu, Y. Liu, L. Shang, X. Liu, and H. Li, "Neural machine translation with reconstruction," in AAAI Conference on Artificial Intelligence, 2017.