Image Captioning And Audio Conversion

Athithiyan G[1], Bharathikannan G[2]
[1]Dept of CSE
[2]Asst Professor Dept of CSE
[1, 2] Sembodai Rukmani Varatharajan Engineering College

Abstract- *Image Sentence Captioning and audio conversion is a phenomenon that describes an image in the form of audio. It is mainly used in applications where one needs information from any particular image in an audio format automatically. This project overcomes limitations by generating entire sentences for describing pictures then changing the text into audio, which can tell detailed, unified stories. It develops a deep model that decomposes both images and sentences into their constituent elements and sleuthing linguistics regions in images with the help of RNN and NLP techniques. It also presents the implementation of LSTM and TTS in Deep Learning techniques as additional features for a good performance.*

Keywords- Deep Learning, LSTM, TTS, RNN, NLP

I. INTRODUCTION

Image captioning aims to describe the object, actions, and details present in an image using NLP. Most image captioning research has focused on single captions, but the descriptive capacity of this form is limited; a single caption can only describe in detail a small aspect of images. Recent work has argued instead for image sentence captioning to generate a sentence describing an image. Compared with single-word captioning, sentence captioning is a relatively new task. The main sentence captioning dataset is the Visual Genome corpus, introduced by Krause et al. When strong single-word captioning models are trained on this dataset, they produce a repetitive word that is unable to describe diverse aspects of images. The generated sentence repeats a slight variant of the same word multiple times, even when beam search was used.

Likewise, different methods are used for sentence generation, they are Long-term Recurrent Convolutional Network: The input can be an image or sequence of images from a video frame. The inputsare given to CNN which recognizes the activities in the image and a vector representation is formed and given to the LSTM model where a word is generated and a caption is obtained. Visual sentence Generation: This method implies giving a coherent and detailed generation of the words. Few semantic regions are detected in an image using the attention model and words are generated one after the other and a sentence is generated.

A recurrent neural network (RNN) is a neural network that is specialized for processing a sequence of data with a timestamp index t from 1 to t. This process will get the max time as t

For a task that involves sequential inputs, such as speech and language, it is often better to use RNNs, Inan NLP problem if you want to predict the next word in a sentence it is important to know the words before it.

The gated recurrent unit (GRU) is a relatively recent development proposed by Cho et al. Similar to the LSTM unit, the GRU has gating units that modulate the flow of information inside the units, however, without having a separate memory cell. Gated Recurrent Unit (GRU) calculates two gates called to update and reset gates which control the flow of information through each of the hidden units. Each hidden state at time-step t is computed using the following equations: Update gate formula, Reset gate formula, new memory formulas, Final memory formula. The update gate is calculated from the current input and the hidden state of the previous time steps. This gate controls how much of portions of new memory and old memory should be combined in the final memory. Likewise, the reset gate is calculated but with a different set of weights. It can control the balance between previous memory and the new input information in the new memory. And TTS is used for the audio conversion of the text.

II. RELATED WORK

Marc'AurelioRanzato, Sumit Chopra, Michael Auli, Wojciech Zaremba had discussed These models are typically trained to predict the next word in the sequence, given the previous words and some context such as the image. However, at test time the model is expected to generate the entire sequence from the scratch. This discrepancy makes generation brittle, as an error may accumulate along the way. We addressed the issue by proposing a novel sequence-level training algorithm that directly optimizes the metric used at test time- BLEU or ROUGE. On three different methods, our

approach outperforms strong baselines for a greedy generation.

Peter Anderson, Xiaodong He, Chris Buehler, Damien, Mark Johnson, Stephen Gould, Lei Zhang had discussed Problems combining image and language understanding such as image captioning and visual question answering continue to inspire considerable research at the boundary of computer vision and natural language processing. In both these tasks, it's often necessary to perform some fine-grained visual processing or even multiple steps of reasoning to generate high-quality outputs. This mechanism improves performance by learning to focus on the regions of the image that are salient and are currently based on deep neural network architectures. A top-down attention mechanism, Faster R-CNN LSTM is used to achieve this.

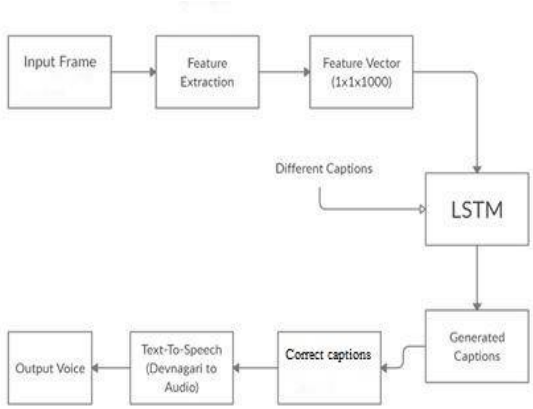
Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, Erkut Erdem provided an in-depth evaluation of the existing image captioning metrics through a series of carefully designed experiments.we explore the utilization of the recently proposed Word Mover's Distance (WMD) document metric for image captioning.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, Eric P. Xing recovered a natural image that usually conveys rich semantic content and can be viewed from a different angle. Existing image description methods are largely restricted by small sets of biased visual sentence annotations, and fail to cover rich underlying semantics.

WaveNet (van den Oord et al., 2016) is a powerful generative model like audio. It works well for TTSbut is slow due to its sample-level autoregressivenature. It requires conditioning on linguisticfeatures from an existing TTS frontend, and not end-to-end: it only replaces the vocoder andacoustic model. Another recently-developed neural network model is DeepVoice, whichreplaces every single component in a typical TTS pipeline with a corresponding neural network. However,each component was independently trained, and it is nontrivial to change the system to train in anend-to-end fashion.

Wang et al. (2016) is the earliest work touching end-to-end TTS using seq2seqwith attention. However, it required a pre-trained hidden Markov model (HMM) aligner to help theseq2seq model learn the alignment. It is hard to tell how much alignment is learned by the seq2seqper second, a few tricks are used to get the model trained, which the author's note hurts prosody.Last, it predicts vocoder parametershence needs a vocoder. Furthermore, the model is trained inphoneme inputs and the experimental results seem to be somewhat limited.

III. ARCHITECTURE



IV. SYSTEM DESIGN

Our existing system can give only a single word as output. We can't able to get an entire sentence through this method.

Disadvantages:

- Not a faster one to predict words.
- It Can't give audio as output.

Proposed System:

The disadvantages of the existing system were overcome by the proposed work.

Advantages:

- Faster than other existing system predictions.
- It can generate audio as output

Disadvantages:

- Does not provide all-time accurate values

V. IMPLEMENTATION

User Interface Design

The user interface has been designed to interact with the user. The elements in UI are,

- **Input Box:** It will accept any size of the image and push it to the model.
- **Output Box:**This will return the captioned text for the particular image.
- **Recorder:** It is used to show the audio as outputfor captions.

Data MiningFor Individual

This module involves the training of the system with the dataset using the Deep Learning algorithm. This involves the following process: Data Collections, Data Cleaning, Data Modeling,and Data Visualization. The analysis is done using the Deep Learning algorithm. NLP is used todevelop the text form image. The attributes have the value of a single sentence as output. Then it will be converted to audio using TTS methods.

- Sensitivity or TPR= TP/P
- Specificity or SPC=TN/N
- Precision Value: PPV = TP/ (TP+FP)
- Accuracy, ACC = (TP+TN) / (TP+FP+FN+TN)
- F1 is harmonic mean of precision and sensitivity.

$$F1 = 2TP / (2TP+FP+FN)$$

The efficient algorithm is chosen based on the above-calculated results.

Finding Vulnerability

In this module, a huge amount of data has been fed into the application to find the text of the image and this text will be converted to audio. This will work from the history of datafrom the dataset.So, this may end increating a wrong output, this must be evaluated and to be reduced in the training part.

Evaluation Metrics

Severaltypes of evaluation methods are usedto measure the quality of the generated captions compared to the ground truth. Each metric appliesits technique for computation and has distinct advantages. The commonly used evaluation are,

- BLEU
- ROUGE
- CIDEr
- METEOR

Outcome

After all, the evaluation input image will be converted to audio. This audio record can be usedfor future purposes.

VI. CONCLUSION AND FUTURE ENHANCEMENT

Image captioning has higher average precision than other methodologies used. The proposed method clearly shows a higher precision in identifying classes such as flowers,beaches, elephants,buildings, mountains, and food compared to the other methods. A class like Dinosaur seems to be better identified in other methodologies. This might be attributed to the lack of images belonging to a similar category in the Flickr8k dataset. This model performs better in image captioning based on precision compared to the previous experiments.

Image captioning has become an important problem in recent days due to the exponential growth of images in social media and the internet. In this report, the various research in image retrieval used in the past and it also highlights the various techniques and methodology used in the research. Hence, complete research in image retrieval making use of the context of the images such as image captioning will facilitate solving this problem in the future. This project can be further enhanced in the future to improve the identification of classes which has a lower precision by training them with more image captioning datasets.

REFERENCES

[1] Reza Hassanzadeh, John H. Phan and May D. Wang, “A Multi-Modal Graph-Based Semi-Supervised Pipeline for Predicting Cancer Survival”, 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).
[2] Saranya P and Satheeskumar B, “A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques”, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 713-719.
[3] Dmitry Ignatov and Andrey Ignatov, “Decision Stream: Cultivating Deep Decision Trees”, 3 Sep 2017 IEEE.
[4] Kelvin KF Tsoi1, Yong-Hong Kuo and Helen M. Meng, “A Data Capturing Platform in the Cloud for Behavioral Analysis Among Smokers An Application Platform for Public Health Research”, 2015 IEEE.
[5] Gonzalez-Alonso P, Vilar R, Lupiañez-Villanueva F, “Meeting Technology and Methodology into Health Big Data Analytics Scenarios”, 2017 IEEE.

[6] B.N. Lakshmi, G.H. Raghunandhan "A Conceptual Overview of Data Mining", 2011 IEEE conference.pp.27-32.
[7] "A Machine Learning-Based Approach for Detection of Alzheimer's Disease Using Analysis of Hippocampus Region from MRI Scan",IEEE 2017
[8] A¨aron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves,Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model forraw audio. arXiv preprint arXiv:1609.03499, 2016.
[9] Liping Chen, Yan Deng, Xi Wang, Frank K Soong, and Lei He. Speech bert embedding for improving prosody in neural tts. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6563–6567. IEEE, 2021.