IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204334 |

Image2speech Using Natural Language Processing(NLP): Automatically Generating Audio Descriptions of Images

Chaitra R [1], Impana J [2], Priyanka M V [3], Venkatesh B S [4], Prof. Raghu B R [5], Prof.Shwetha G [6]

U.G. Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India[1]

U.G. Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India[2]

U.G. Student Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India[3]

U.G. Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India[4]

Assistant Professor, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India[5]

Assistant Professor, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India[6]

ABSTRACT: This project proposes a new task for Artificial Intelligence.The Image2speech task generates a spoken description of an image. We present baseline experiments in which the neural net used is a sequence-to-sequence model with attention, and the speech synthesizer is clustergen. Speech is generated from four different types of segmentations: two that require a language with known orthography (words and first-language phones), and two that do not (pseudo-phones and second-language phones). BLEU scores and token error rates indicate that the task can be performed with better than chance accuracy. Informal perusal of the output (phone strings, word strings, and synthesized audio) suggests that the audio contains complete, intelligible words organized into intelligible sentences, and that the most salient errors are caused by mis-recognition of objects and actions in the image.

KEYWORDS: Artificial Intelligence,Image2speech, Natural Language Processing (NLP),, Convolutional Neural Network (CNN).

I. INTRODUCTION

This project proposes a new task for artificial intelligence: the generation of a spoken description of an image. The automatic generation of text is the topic of natural language processing (NLP), whereas the analysis of images is the topic of the field of computer vision. In both fields, great advances have been made on these separate topics, and recently they have been combined into a new research field: img2txt . However, many of the world's languages do not have a written form , therefore many people do not have access to these and other speech and NLP technologies. In this work, we propose a new research task: Image2speech, which is similar to img2txt, but can reach people whose language does not have a natural or easily used written form. An Image2speech system should generate a spoken description of an image directly, with first generating text.

Experiments reported in this paper convert image feature vectors into speech unit sequences. In order to implement this pipeline, four types of standard open-source software toolkits are used. First, the Visual Geometry Group (VGG16) visual object recognizer converts each image into a sequence of feature vectors. Second, the Extensible Neural Machine

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204334 |

Translation toolkit (XNMT)machine translation toolkit accepts image feature vectors as inputs, and generates speech units as output. Third, the ClusterGen speech synthesis toolkit generates audio from each speech unit sequence. Fourth, in order to train a synthetic speech voice, Clustergen needs a corpus of audio files, each of which is transcribed using some type of discrete symbolic units; automatic speech recognition (ASR) systems based on Kaldi and Eesen perform this transcription.

The complete image2speech system is trained using a corpus of (image, description) pairs, where each description is an audio file. Four different types of speech units are tested, distinguished by the type of technology used to segment the audio training data. Two types of unit sequences, Words and L1-Phones (first-language phones), are generated using a same language ASR, and would therefore never be applicable to a language without orthography, but they provide us with an upper bound performance on the Image2speech.

Two other unit sequences, L2-Phones and Pseudo-Phones, are generated without transcribed same-language speech, and would therefore be applicable even in a language lacking orthography. L2- phones (second-language phones) are generated by an ASR that has been trained in some other language. Pseudo-phones are generated by an unsupervised acoustic unit discovery system.

II. LITERATURE SURVEY

In 2015, Harwath and Glass proposed a data retrieval system in which speech files are used to retrieve corresponding images from a large image database, and vice versa. In order to make their proposal possible, they hired crowd workers on Mechanical Turk to read aloud the 40,000 captions from the Flickr8k corpus. The resulting set of 40,000 spoken captions is distributed as the Flicker-Audio corpus.

Imagenet is an image database organized according to the WordNet noun hierarchy. ImageNet currently has 14m images, provided as examples of 22k nouns. The ILSVRC (Imagenet Large Scale Visual Recognition Challenge) has been held annually since 2010. The best single-network solution in ILSVRC 2014 Sub-task 2a, "Classification + localization with provided training data," was a 13-layer convolutional neural network (CNN) ; Implementations in TensorFlow (used in this paper) and Keras are now redistributed as the VGG16 network. VGG16 is a 13-layer CNN, followed by a two-layer fully-connected network (FCN). The last convolutional layer is composed of 512 channels, each of which is a 14×14 image; it is useful to interpret this layer as a set of 14×14 = 196 feature vectors of dimension 512. Each feature vector is the nonlinear transformation of a 40 × 40-pixel subimage, which is to say, about 3% of the original 224 × 224 input-image.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015.

Imagenet large scale visual recognition challenge. IJCV, 115(3):211–252.Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems.

III. METHODOLOGY

Image2speech

Fig. 1 gives an overview of experimental methods used in this project. Image2speech models were trained using (image,audio) pairs drawn from the Flickr8k, MSCOCO, FlickerAudio, and SPEECH-COCO corpora. Each image is represented as a sequence of 196 vectors, each of dimension 512, created from the last convolutional layer of the VGG16 network. Audio files are converted to units via Kaldi forced alignment (Words and L1- phones) or via Eesen or AMDTK phone recognition (L2-phones and Pseudo-phones). XNMT then learns to convert a sequence of image feature vectors into a sequence of speech units, while Clustergen learns to convert speech units into audio.

The image2speech model learned by XNMT is a sequenceto-sequence model, composed of an encoder, an attender, and a decoder. The encoder is a one-layer bidirectional LSTM (implemented using XNMT's PyramidalLSTM model), with a 128- dimensional state vector. The attender is a three-layer perceptron, implemented using XNMT's StandardAttender model. For each combination of an input LSTM state vector and an output LSTM state vector (128 dimensions each), the attender uses a three-layer perceptron (two hidden layers of 128 nodes each) to compute a

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204334 |

similarity score. The decoder is another three layer perceptron (1024 nodes per hidden layer), which views an input created as the attention-weighted summation of all input LSTM state vectors, concatenated to the state vector of the output LSTM. The output of the decoder is a softmax with a number of output nodes equal to the size of the speech unit vocabulary.

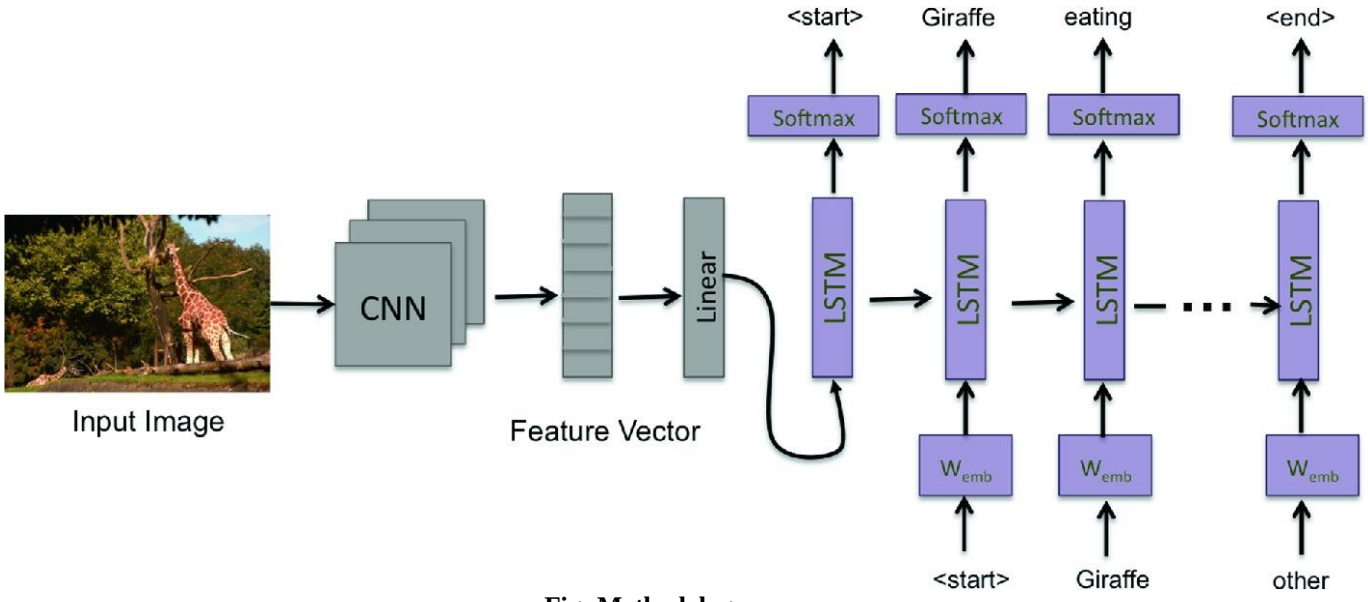

Fig :Methodology

Flickr8k dataset :

A new benchmark collection for sentence-based image description and search, consisting of 8,000 images that are each paired with five different captions which provide clear descriptions of the salient entities and events. The images were chosen from six different Flickr groups, and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations.

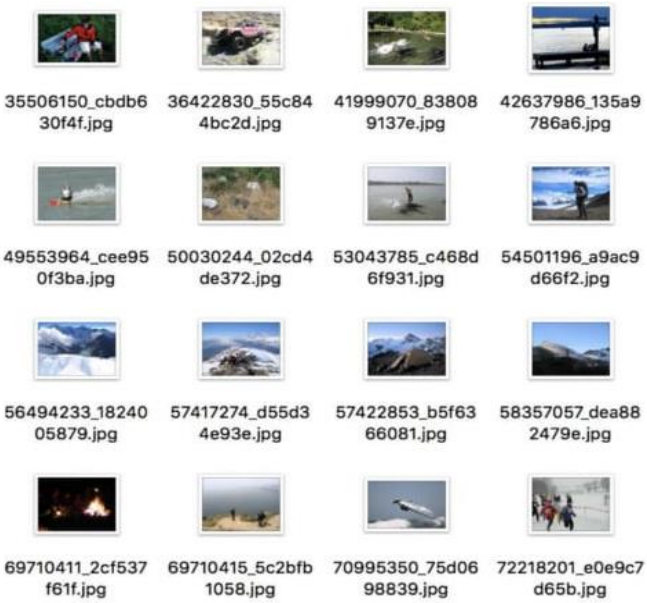

Fig :Flickr8k dataset

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204334 |

IV. EXPERIMENTAL RESULT

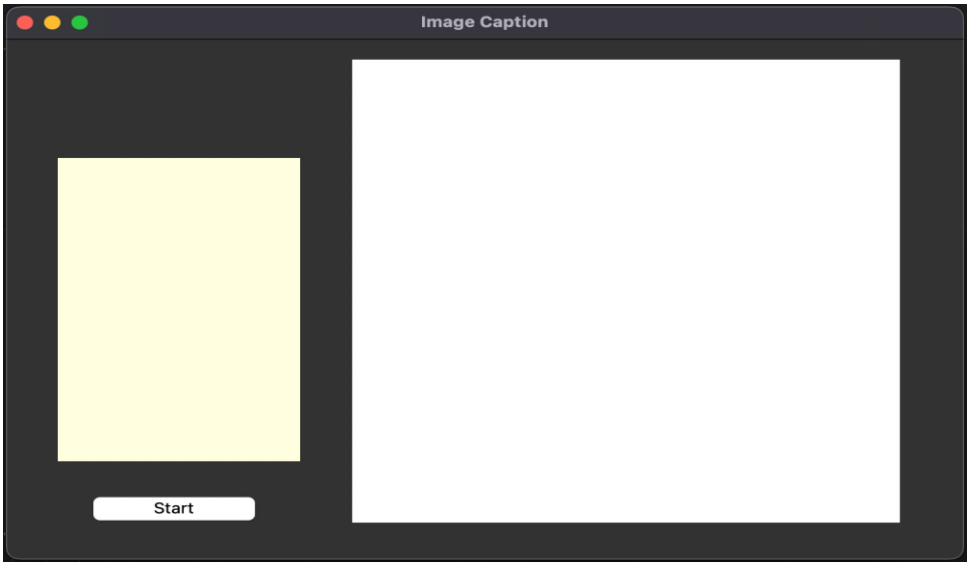

fig.1:User Interface Desgin of Image2Speech

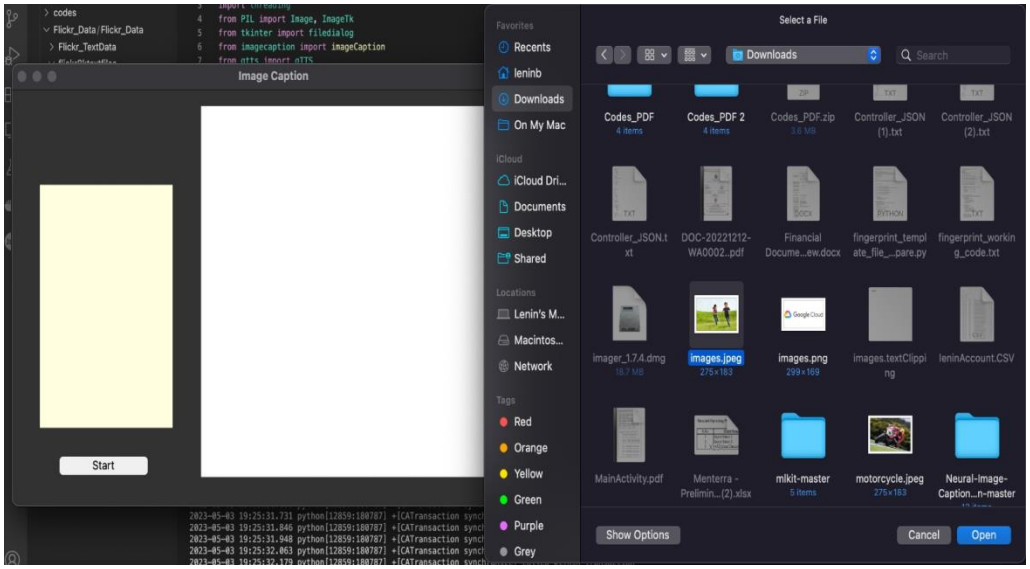

fig 2:Selecting an image from Flickr8k dataset.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204334 |

fig3: Generated Image Caption with Text Description.

fig 4:Generated Image Caption with Audio Description

V. CONCLUSION

This paper proposes a new task for artificial intelligence: image2speech, the task of generating spoken descriptions of input images, with intermediate text. Image2speech is trained using a database of paired images and audio descriptions. Experimental results are presented using the Flicker-Audio and SPEECH-COCO corpora. Measured UER scores are better than chance, but less than perfect. Informal perusal of results shows that image2speech is able to generate intelligible words, and to sequence them into intelligible sentences.

REFERENCES

[1] J-Y Pan, H-J Yang, P Duygulu, and C Faloutsos, "Automatic image captioning," in Proc. IEEE Internat. Conf. Multimedia and Expo (ICME), 2004.
[2] G Adda, S Stuker, M Adda-Decker, O Ambouroue, ¨ L Besacier, D Blachon, M Bonneau-Maynard, P Godard, F Hamlaoui, D Idiatov, G-N Kouarata, L Lamel, E-M Makasso, A Rialland, M Van de Velde, F Yvon, and S Zerbian, "Breaking the unwritten language barrier: The BULB project," in Proceedings of the SLTU-2016 5th Workshop on Spoken Language Technologies for Underresourced languages, 2016.
[3] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image classification", https://arxiv.org/abs/1409.1556, 2014, Accessed: 2017- 09-14.
[4] D Frossard, "Vgg16 in tensorflow," https://www.cs.toronto.edu/ frossard/post/vgg16/, 2016, Accessed: 2017-09-14.
[5] G Neubig, "eXtensible Neural Machine Translation," https://github.com/neulab/xnmt, 2017, Accessed: 2017- 09-14.
[6] AW Black, "CLUSTERGEN: A statistical parametric speech synthesizer using trajectory modeling," in Proc. Internat. Conf. Spoken Language Process. (ICSLP), 2006, pp. 1762–1765.
[7] D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlicek, Y Qian, P Schwarz,

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204334 |

J Silovsky, G Stemmer, and K Vesely, "The Kaldi speech recognition toolkit," in Proc. of ASRU, 2011.
[8] Y Miao, M Gowayyed, and F Metze, "EESEN: Endto-end speech recognition using deep RNN models and WFST-based decoding," in IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015.
[9] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei, "Construction and Analysis of a Large Scale Image Ontology," in Vision Sciences Society, 2009.
[10] C Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details