Visual Assistance For Visually Impaired People Using Image Caption and Text To Speech

Venkataramana N[1,a)]Nagesh P[2,b)] Prabha B [3,c)]

[1,2,3,4,5]*Department of Computer Science and Engineering,Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302*

[a)]*ramana@kluniversity.in*
[b)]*pnagesh@kluniversity.in*
[d)]*jemi.prabha @kluniversity.in*

DOI : 10.48047/IJFANS/11/S6/039

Abstract .Picture captioning has recently become a new difficult challenge that gathered everyone interest, which is being able to automatically define an image's content with properly formatted text English sentences. it can make a great impact by assisting people who are visually impaired better recognitionof theircircumstances. By taking the images of surrounding environment then make use of these photos to generate captions that can be read out visual amplification impaired, so that they can get a better sense of what's going on around them. In this p a per To extract features, we used a combination of convolutional neural networks of the images and then LSTM was used (Long short-term memory) to generate text from these features. The obtained text is then converted into speech so that it can be read out. Our model generates highly descriptive captions that canpotentially greatly improve the lives of visually impaired people.

Keywords:Xception**,**,LSTM,intellegence

Introduction

Computer vision has made considerable advances in the image processing field in recent years, such as image recognition and object detection. It is now possible to automate the process create one or more sentences to help you understand what you're doing of an image, thanks to technological advancements in image recognition and object detection. This is referred to as Image Captioning. Automatically generating full and natural image descriptions has a wide variety of applications[1], such as titles for news images, descriptions for medical images, text-based image retrieval, details for blind users, and human-robot interaction. These image captioning applications are useful for the purpose of both theoretical and practical research. As a result, image captioning has become a more difficult but important job in the age of artificial intelligence [2].

When given a new image, an image captioning algorithm should produce a semantic description of the image. In Fig. 1, for example, the input image includes individuals, motor cycle, man, and mountain [3]. A sentence at the bottom describes the image's content about the objects that appear in the image, the action, and the scenes are all identified in this sentence.

Humans can easily comprehend image content and articulate it in natural language sentences according to uniqueneeds for image captioning; however, computers need the integrated use of image processing, computer vision, natural language processing, and other major areas of research findings for this mission[4]. The goal of image captioning is to build a model that can be used in its entirety image data in order to produce richer, more

human-like image descriptions. The meaningful definition generation process

Figure: input image includes individuals, motor cycle, man, and mountain

in high level image semantics necessitates not only object or scene recognition in the image, but also the ability toanalyze their states, comprehend their relationships, and produce a semantically and syntactically correct sentence. It's still a mystery how the brain comprehends an image and organises the visual data into a caption. Image captioning necessitates a thorough understanding of the environment and which elements are essential to the whole[5]. We used a combination of convolutional neural networks to extract features, and then created text from these features using LSTM (Long short-term memory). And we are using GttsAPI for converting captions to audio file. In Python, there are many APIs for converting text tovoice. The Textto Speech by Google API, also known as the gTTS API, is one of these APIs. gTTSis a straightforward method that transforms entered text into audio that can be saved as an mp3 file. English, Hindi, Tamil, French, German, and several other languages are supported bythegTTS API.

CNN-RNN based framework

Image data is mapped to an output variable using Convolutional Neural Networks. They've proven to be so efficient that they're now the method of choice for any form of prediction problem involving image data as an input. RNNs, or recurrent neural networks, were developed to solve sequence prediction problems.

One-to-many, many-to-one, and many-to-many are examples of sequence prediction problems. The most effective RNNs are LSTM networks, which allow us to encapsulate a larger sequence of words or sentences for prediction[6-7].

The blending of various types of networks into hybrid models creates one of the most enthralling and realistic neural models. Consider the job of creating image captions. There is an input image and an output image sequence, which is the caption for the input image in this case. The input image is incomprehensible to LSTM or any other sequence prediction model. Since they aren't designed to deal with such inputs, we can't explicitly input the RGB image tensor. The LSTM struggles to model input with spatial structure, such as photographs.

Using the deep CNN architecture, we can extract features from the picture, which are then used to generate the caption by the LSTM architecture. The CNN LSTM model was created specifically for sequence prediction problems with spatial inputs, such as images or videos. Convolutional Neural Network (CNN) layers for feature extraction on input data are combined with LSTMs for sequence prediction on the feature vectors in this architecture[8]. In a nutshell, CNN LSTMs are a type of model that is both spatially and temporally deep and is the intersection of computer vision and natural language processing has been discovered. These models have a lot of promise, and they're becoming more popular for tasks like text classification and video conversion.

Proposed method.

Proposed Methodology and Architecture

We will use a combination of CNN and LSTM to incorporate the caption generator in this suggested approach (Long short-term memory). The image features will first be extracted using Xception, a pre- trained CNN model trained on the ImageNet dataset, and then the features will be fed into the LSTM model, which has been trained on the caption of the flicker8k dataset, to recall the memory and produce the image captions[9].

We will use a combination of CNN and LSTM to incorporate the caption generator in this suggested approach (Long short-term memory). The image features will be extracted first using Xception, a pre- trained CNN model trained on the ImageNet dataset, and then fed into the LSTM model. which is trained by the caption of flicker8k dataset so that it will recall the memory and will be responsible for generating the imagecaptions.

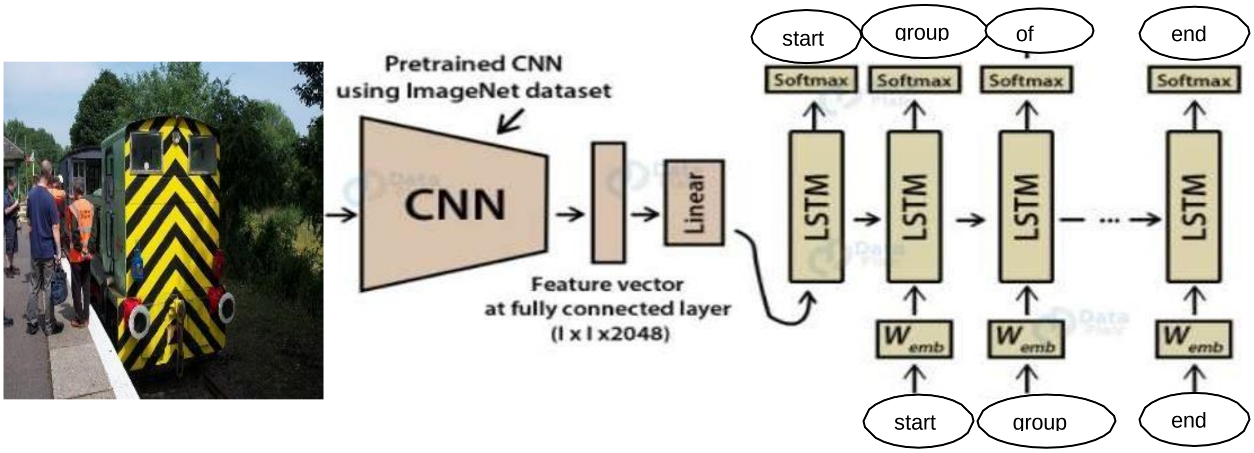


Figure: a. Block Diagram

b. Algorithm:

Step 1 : The image is given as input.

Step 2 : The image is passed through xceptionmodels.which collects the features.

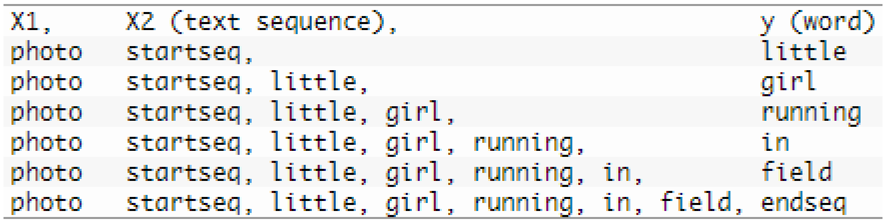
Step 3 : Features passed through LSTM.

Step 4 : Feature vectors of image.

Step 5 : Caption generated converted to speech.

The process begins by passing the image through the pre-trained model Xception as an input. The features of the layer before the last layer are generated by the Xception model (n-1).

After collecting the features, it passes through the LSTM model which recalls its memory of training from intheflicke8k dataset's captions Itproduces the caption ina word-by-word sequence,asseeninthe Diagrambelow..



```
X1,     X2 (text sequence),                         y (word)
photo   startseq,                                   little
photo   startseq, little,                           girl
photo   startseq, little, girl,                     running
photo   startseq, little, girl, running,            in
photo   startseq, little, girl, running, in,        field
photo   startseq, little, girl, running, in, field, endseq
```

Figure: Word by Word Sequence flow.

The created words will be concatenated and recursively given as input to create a caption for an image when the model is used to generate descriptions.

ExperimentalInvestigations

XceptionModel

Xception Inspired by Google's Inception model. Xception is based on an 'extreme' interpretation of the Inception model. The Xception architecture is a linear stack of depth wise separable convolution layers with residual connections, Simple and modular architecture.

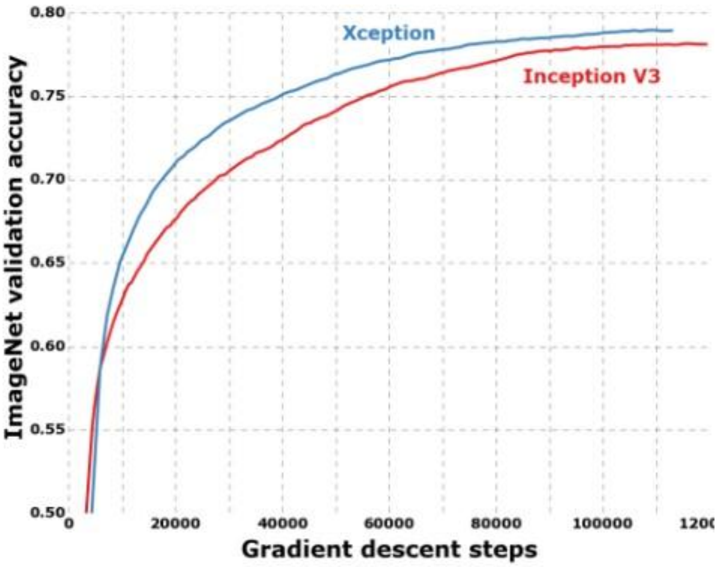


Figure: Xception and Inception Graph

“Depthwise Separable Convolution Regular Convolutions:

- look at both channel & spatial correlations simultaneously

Depthwise separable convolution:

- look at channel & spatial correlations independently in successivesteps
- spatial convolution: 3x3 convolutions for eachchannel
- depth wise convolution: 1x1 convolutions on concatenatedchannels

Example: take 3x3 convolutional layer on 16 input channels and 32 output channels.

- regular convolution: 16x32x3x3 = 4608parameters
- depthwise separable convolution: (spatial conv + depthwise conv) = (16x3x3 + 16x32x1x1) =656 parameters
- greatly reduced parametercountmoreefficientcomplexitymaintains cross-channelfeatures

The data first passes through the entry flow, then eight times through the middle flow, and finally through the exit flow. Batch normalisationis extended on both Convolution and Separable Convolution layers. Xception architecture has overperformed VGG-16, ResNet and Inception V3 in most classicalclassificationchallenges.

Xception is an efficient architecture that relies on two main points:

- DepthwiseSeparableConvolution

- Shortcuts between Convolution blocks as inResNet”

Depth wise Separable Convolution

Depth wise Separable Convolutions are alternatives to classical convolutions that are supposed to be much more efficient in terms of computation time.

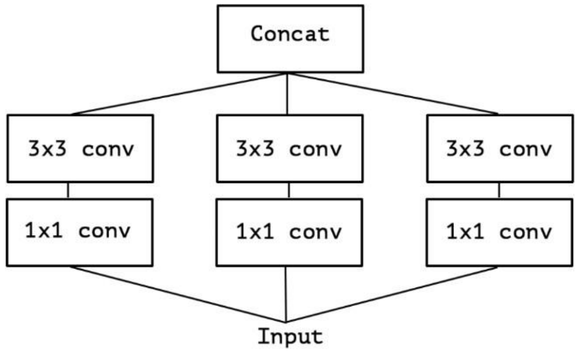


Figure Inception Model

- “Fundamental hypothesis: The spatial and cross-channel associations are sufficientlydecoupled.
- First looks at cross channel correlations via a set of 1x1convolutions.
- then acts as a “multi-level feature extractor” by computing 1×1, 3×3, and 5×5convolutions

Output feature maps are stacked along the channeldimension

“extreme” version of Inception module:

- first use a 1x1 convolution to map cross-channelcorrelationsthen separately map the spatial correlations of every output channel (instead of just 3-4partitions)
- Similar to depth wiseseparableconvolution”

Xceptionarchitecture

The design of a convolutional neural network is entirely made up of depthwise separable convolutionlayers.

The basic hypothesis is that cross-channel and spatial correlation mapping can be fullydecoupled.

The network's feature extraction base is made up of 36 convolutionallayers[10].

Except for the first and last modules, which have linear residual connections around them, the structure is divided into 14modules.

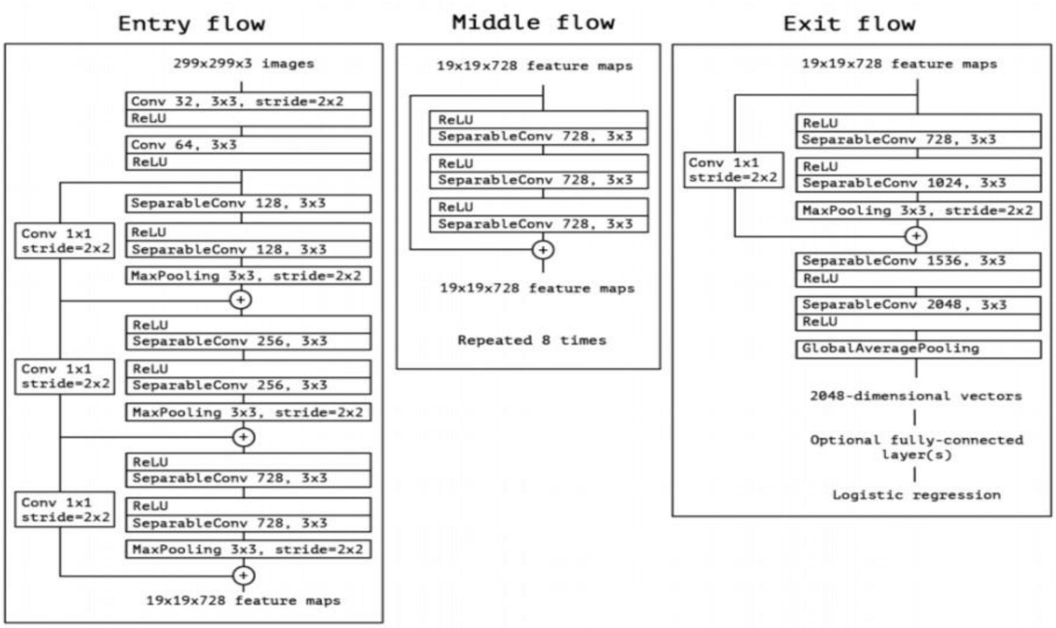
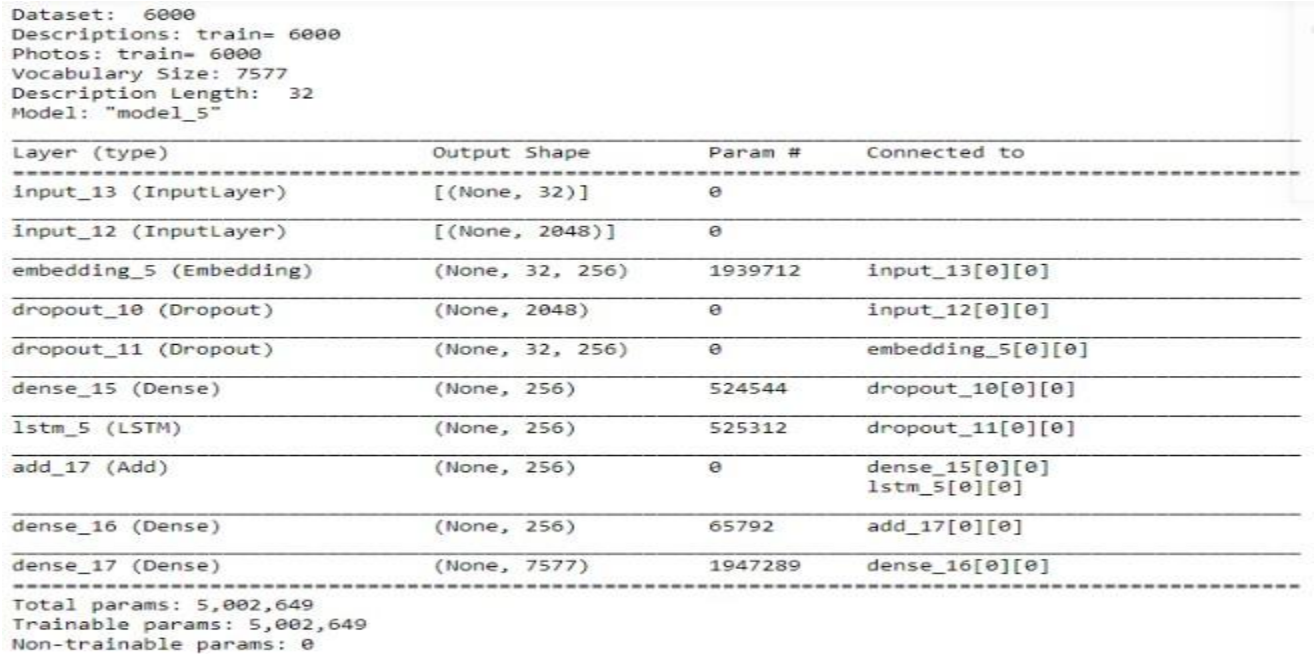


Figure: Xceptionarchitecture

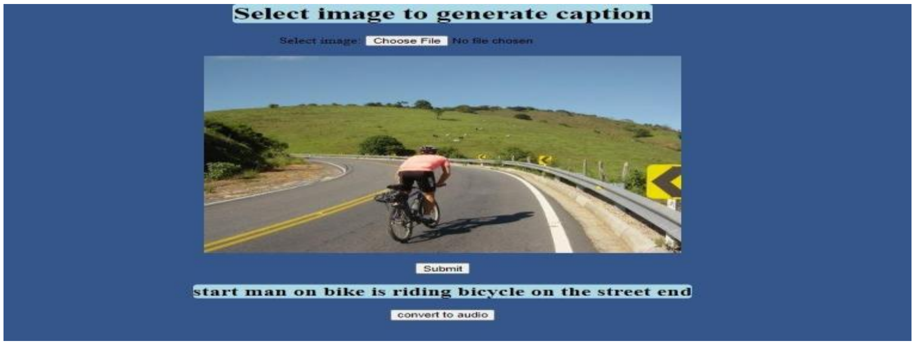


```
Dataset:  6000
Descriptions: train= 6000
Photos: train= 6000
Vocabulary Size: 7577
Description Length:  32
Model: "model_5"
______________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==============================================================================================
input_13 (InputLayer)           [(None, 32)]         0
______________________________________________________________________________________________
input_12 (InputLayer)           [(None, 2048)]       0
______________________________________________________________________________________________
embedding_5 (Embedding)         (None, 32, 256)      1939712     input_13[0][0]
______________________________________________________________________________________________
dropout_10 (Dropout)            (None, 2048)         0           input_12[0][0]
______________________________________________________________________________________________
dropout_11 (Dropout)            (None, 32, 256)      0           embedding_5[0][0]
______________________________________________________________________________________________
dense_15 (Dense)                (None, 256)          524544      dropout_10[0][0]
______________________________________________________________________________________________
lstm_5 (LSTM)                   (None, 256)          525312      dropout_11[0][0]
______________________________________________________________________________________________
add_17 (Add)                    (None, 256)          0           dense_15[0][0]
                                                                 lstm_5[0][0]
______________________________________________________________________________________________
dense_16 (Dense)                (None, 256)          65792       add_17[0][0]
______________________________________________________________________________________________
dense_17 (Dense)                (None, 7577)         1947289     dense_16[0][0]
==============================================================================================
Total params: 5,002,649
Trainable params: 5,002,649
Non-trainable params: 0
```

Figure:ExperimentalResults

FIGURE-1:

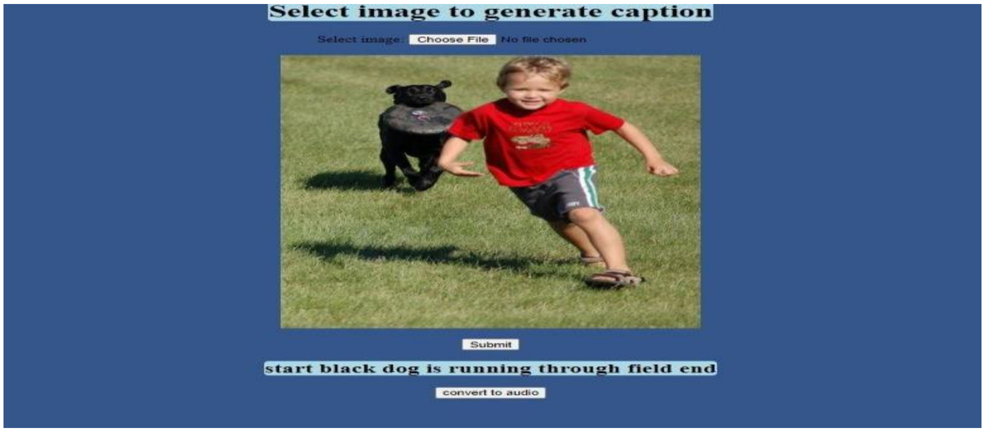
The preceding demonstrates that the man is riding a bicycle on the sidewalk which fairly a correct prediction. But it may be not a street although it mostly predicted the correct caption.



Select image to generate caption

start man on bike is riding bicycle on the street end

FIGURE-2:



Select image to generate caption

start black dog is running through field end

In the above prediction it predicted that black dog is running through the grass but it is missed the boy but it predicted the dog running through the grass correctly, so we can increase the accuracy by training on more images and captions data.

4.2 Discussion ofResults:

Table: accuracy of Results.

Model	Top-5 Accuracy	Parameters	Depth
VGG16	0.901	138,357,544	23
InceptionV3	0.937	23,851,784	159
ResNet50	0.921	25,636,712	
Xception	0.945	22,910,480	126
InceptionResNetV2	0.953	55,873,736	572
ResNeXt50	0.938	25,097,128	

Accuracy of the CNN model trained on different datasets.Xception outperforms VGGNet , ResNet , and Inception-v3. Xception has better accuracy compared with Inception-v3 along the gradient descent steps. Xception model is trained on image net dataset with 22,910,480 parameters and 126 layers depth network..Xceptions shows an accuracy of 0.945 which overcomes all the previous models accuracy.

Conclusion:

For this experimental results shows the accuracy of the proposed system it definitely helping to blind people to know what is there in surroundings like a sighted people.by using this system blind people can feel like normal people to identify and feel like normal people what is there around them with voice assistance. When consider into time complexity its take very less time to identify and give voice assistance to blind people. The future scope of this proposed system it can also be embedded to driver less vehicle's to identify the articles around the vehicles while driving.

References:

1. "Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge" BinqiangWang ,Xiangtao Zheng , Member, IEEE, Bo Qu, and Xiaoqiang Lu , Senior Member, IEEE 2020.
2. "Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning" Pierre Dognin∗, Igor Melnyk∗ , Youssef Mroueh∗,† , Inkit Padhi∗ , Mattia Rigotti∗ , Jarret Ross∗ , Yair Schiff∗ , RichardA. Young, and Brian Belgodere 21 Dec 2020
3. "Topic-Oriented Image Captioning Based on Order-Embedding." Niange Yu, Student Member, IEEE,Xiaolin Hu , Senior Member, IEEE, Binheng Song, Jian Yang, and Jianwei Zhang . 6, JUNE 2019.
4. Multisource Image Fusion Method Using Support Value Transform, Sheng Zheng, Wen-Zhong Shi, Jian Liu, Guang-Xi Zhu, and Jin-Wen Tian, IEEE Transactions On Image Processing, Vol. 16, No. 7, July 2007.
5. S. G. Nikolov and D. R. Bull et al., "Image fusion using A 3-D wavelet transform," in Inst. Elect. Eng. Conf. Pub., vol. 1, 1999, pp. 235–239.
6. "Non-Autoregressive Text-to-Speech Synthesis based on Very Deep VAE with Residual Attention"Chaw Su Thu Thu , Theingi Zin, march -2014.s

7. Silva, I.; Moody, G.B.; Scott, D.J.; Celi, L.A.; Mark, R.G. Predicting in-hospital mortality of ICU patients: The physionet/computing in cardiology challenge 2012. Computing in Cardiology. Krakow. 2012. https://pubmed.ncbi.nlm.nih.gov/24678516
8. Subramanian, B.; Saravanan, V.; Nayak, R. K.; Gunasekaran, T.; & Hariprasath, S. Diabetic Retinopathy–Feature Extraction and Classification using Adaptive Super Pixel Algorithm. Int J Eng Adv Technol. 2019, 9, 618-627. https://doi.org/10.35940/ijeat.B2656.129219
9. Thongprayoon, C.; Cheungpasitporn, W.; Mao, M.A.; Sakhuja, A.; Erickson, S.B. Admission Hyperphosphatemia Increases the Risk of Acute Kidney Injury in Hospitalized Patients. J. Nephrol. 2018, 31, 241–247. https://doi.org/10.1007/s40620-017-0442-6
10. Transl. Med. 2019, 7, 55. https://doi.org/10.21037/atm.2018.06.50
11. Vairavan, S.; Eshelman, L.; Haider, S.; Flower, A.; & Seiver, A. Prediction of mortality in an intensive care unit using logistic regression and a hidden Markov model. In 2012 Computing in Cardiology. 2012, pp. 393-396. IEEE. https://ieeexplore.ieee.org/abstract/document/6420413
12. Zhou, T.; Chung, F. L.; & Wang, S. Deep TSK fuzzy classifier with stacked generalization and triplely concise interpretability guarantee for large data. IEEE Transactions on Fuzzy Systems. 2016, 25(5), pp. 1207-1221. https://doi.org/10.1109/TFUZZ.2016.2604003.