

# Image summarization using hybrid CNN- LSTM networks and corresponding Speech Synthesis using Mel Spectrogram Predictions

Roshan Varghese<sup>1</sup>, Durga Srilekha Regulagadda<sup>2</sup>, S Yoshitha Manavi<sup>3</sup>  
<sup>1,2,3</sup>Department of Computer Engineering, CHRIST Deemed to be University, India

## ABSTRACT

Be it for assistance, or for a more automated entity in a device, image-to-text-to-speech is an up and coming trend in today's computerized world. In this paper, we deal with summarising or captioning, an image that is being fed to the system. Moving it a step further, the text will then be synthesized into speech. Having various applications for it, such as assistance for physically handicapped to captioning out news headlines, this was considered as one of the promising fields of research and experimentation. This paper attempts to approach this problem with a hybrid CNN-LSTM network model for the image to text conversion, and a mel spectrogram combined with a wavenet generator to help tackle the text-to-speech section.

## 1. INTRODUCTION

With recent trends in data, it is exceptionally obvious that a greater amount of data will keep on being produced over the course of years. It is suspected that our limit to provide services or assistance to customers will be restricted by the type of analysis and knowledge that we can obtain or extract from the data. Images constitute a fair share of information in the large form of media that is used for communication. While summarization of videos and events have been of late enthusiasm to the computer vision and multimedia research community, recent advances in the field of optimization, especially deep learning have shown significant improvements in video summarization. However, Image summarization, although being an important task, continues to be eluded because of the intrinsic difficulties and its disparities from video summarization.

Unlike the text and video summarizers, in the case of the collection of images, there is no temporal sequence between two images to be exploited by the network which has resulted in the problem being esoteric in nature. To remedy this, the following methodology plans to present a model that generates natural language descriptions of images and their surroundings using a combination of Convolutional Neural Networks and LSTM (Long-Short-Term Memory). The synthesis of speech from the given caption is achieved using the a mel spectrogram combined with a wavenet generator. Our approach leverages the datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data.

## 2. LITERATURE REVIEW

Our core visual processing module is a Convolutional Neural Network (CNN) [16, 17], which has emerged as a powerful model for visual recognition tasks. The first application of these models to dense prediction tasks was introduced in R-CNN [18], where each region of interest was processed independently. Further work has focused on processing all regions with only a single forward pass of the CNN, and on eliminating explicit region proposal methods by directly predicting the bounding boxes either in the image coordinate system [19] or in a fully convolutional [20]. The paper by Farhadi et al. investigates methods to generate short descriptive sentences from images. Their contributions include the introduction of a dataset to study the problem and the introduction of a representation intermediate between images and sentences. They also described a discriminative approach that produces very good results at sentence annotation[3]

Several early works [21][22], on image captioning, combine object and scene recognition with a template or tree-based approach to generate captions. Such sentences are typically simple and are easily distinguished

from more affluent human-generated descriptions., [23] address this by composing new sentences from existing caption fragments which, though more human-like, are not necessarily accurate or correct. Summarization will also help in an efficient display and representation of relevant data in multiple industries and will be of value for the current internet and e-commerce industry. Also, with the amount of legal documentation being done in current form summarization is to definitely help people with no background of law to understand the cause and effects of things stated in the clauses. It will also help the user augment his understanding of the language tools being used in legal documents to represent or justify situations as described by Singh et al.[8]

The work done by Sol et al. implements a generative CNN-LSTM model that beats human baselines by 2.7 BLEU-4 points and is close to matching (3.8 CIDEr points lower) the current state of the art. Experiments on the MSCOCO dataset show that it generates sensible and accurate captions in a majority of cases, and hyperparameter tuning using dropout and number of LSTM layers allows us to alleviate the effects of overfitting. Moreover, Soh et al. have used deep convolutional neural networks to generate a vectorized representation of an image that can be fed into a Long-Short-Term Memory (LSTM) network, which generates captions. This model is used as a reference for our summarizer.[1]. Convolutional networks have recently enjoyed great success in large-scale image and video recognition which has become possible due to the large public image repositories, such as ImageNet, and high-performance computing systems, such as GPUs or large-scale distributed clusters as mentioned in the works of Krizhevsky et al, Dean et al, Deng et al. [5][6][7]. CNN's are hierarchical neural networks whose convolutional layers alternate with subsampling layers, reminiscent of simple and complex cells in the primary visual cortex as described by Wurtz R. H et al. [9]. Donahue et al. have shown that convolutional networks with recurrent units are generally applicable to visual time-series modeling, and argued that in visual tasks where static or flat temporal models have previously been employed, LSTM style RNNs can provide significant improvement when ample training data are available to learn or refine the representation. As mentioned by Donahue et al., the resulting model can be trained end-to-end on large-scale image and text datasets, and even with modest training, it provides competitive generation results compared to existing methods[2]

A deep convolutional neural network is used to create a semantic representation of an image, which will be decoded using an LSTM network. In an unrolled LSTM network for the CNN-LSTM model, all LSTMs share the same parameters. The vectorized image representation is fed into the network, followed by a special start of sentence token.[1]. In their work Simonyan et al. investigated the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Their main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small ( $3 \times 3$ ) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers.[4]

### **3. RESEARCH METHOD**

For this model, as mentioned before, the summarization of images is done using a CNN - LSTM hybrid model. We aim to use CNN to approach the problem of image classification and mutate it with LSTM for the summarization of the then classified images. CNN model has been chosen mainly due to its popularity for simple and easy working. Training a CNN model is more efficient[1] than compared to the other neural networks. CNN's are meant to work on computer vision tasks and object detection in the image data space.

Based on the training set that was fed to the CNN network, the algorithm learned the object and the class it belongs to, which can then help predict the class in future inputs of the image. It can also measure how accurate its predictions are. This process introduced multiple challenges during its due course of working, including scale variation, viewpoint variation, intra-class variation, et cetera. Since CNN is not a fully connected network and hence uses fewer parameters by using the same parameters many times. The way CNN works is, it breaks the image into smaller working areas, while a fully connected network uses the entire image as a single entity. This approach is advantageous for two reasons - it is more accurate than its counterparts with the same number of working parameters, and since there are lesser parameters, it is faster. When a CNN model is trained to classify an image, it searches for the features at their base level, such as curvatures or boundaries of the object. Figure 1 depicts the basic working of a CNN model. One of the most famous CNN models was developed in the year 2014 called GoogLeNet, which received an error rate lower than 7% [11]

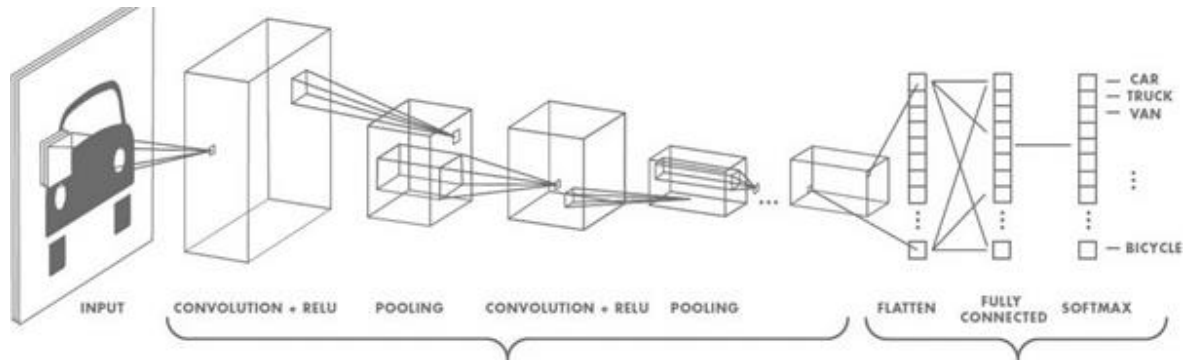


Figure 1: Depiction of the working of a CNN for image classification.

The next step is to ensure the network has a memory that it can use for future predictions and classifications. Since most traditional networks are not capable of having a usable memory[12], recurrent neural networks are implemented. These networks have loops in them which allow information to persist for future use. In Figure 2 below, a chunk of the neural network, A, looks at some input  $x_t$  and outputs a value  $h_t$ . A loop allows information to be passed from one step of the network to the next. Figure 2 below depicts what an RNN would look like if the loops were unrolled.

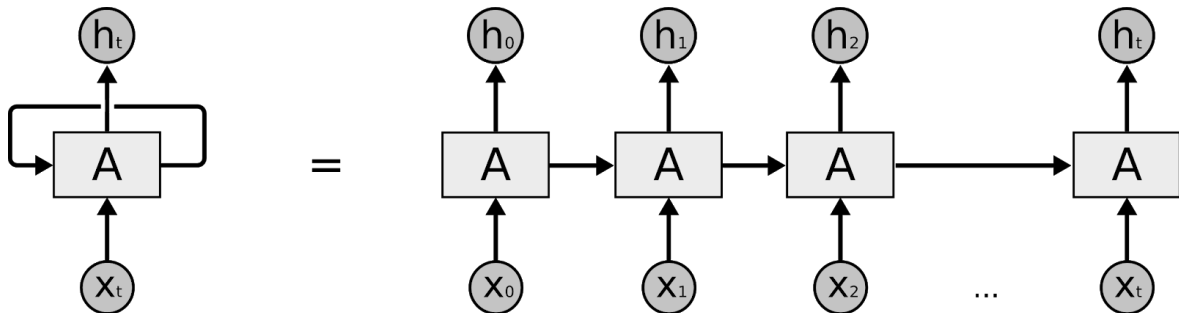


Figure 2: An unrolled recurrent neural network.

In the last few years, RNNs have been applied to various problems such as speech recognition, language modeling, translation, etc[20]. An essential part of RNNs is the use of LSTMs, a special kind of RNNs which is practically a much more efficient version than regular RNNs. Long Short Term Memory networks are capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997) and were refined and modified by many people after that to address various problems. The reason why LSTMs are so efficient is that they are explicitly designed to avoid the long-term dependency problem. While long term dependency is taught to the other neural networks, this algorithm has a long term memory as default. While regular RNNs have a simple single-layered structure, an LSTM algorithm will have multiple layers.

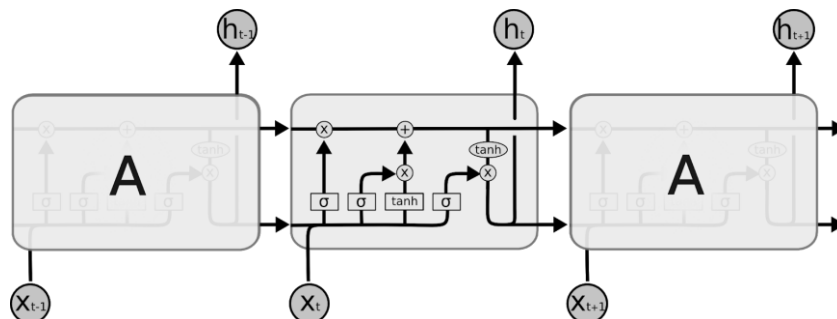


Figure 3: A LSTM neural network.

In the above diagram, i.e., Figure 3, each line carries an entire vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations; the yellow boxes are learned neural network layers. Lines merging denote concatenation, and the forking lines denote its content being copied into different locations. The basis of the model used is neatly depicted in Figure 4.

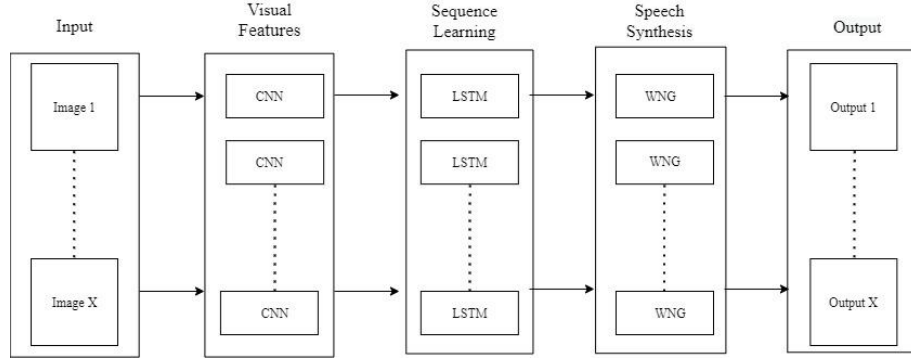


Figure 4: Proposed Diagram

The text to the speech conversion is implemented using a 2 step architecture[24] which comprises of sequence-to-sequence feature prediction network that maps character embeddings to mel-scale spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize time-domain waveforms from those spectrograms.

To produce the magnitude spectrograms from a sequence of characters, the architecture model simplifies the traditional speech synthesis pipeline by replacing the production of linguistic and acoustic features with a single neural network trained from data alone. Griffin-Lim algorithm can be used for further phase estimation to vocode the resulting magnitude spectrums produced.

The WaveNet model produces audio quality that is almost as perfect as that of real human speech and is already used in text to speech synthesis systems mostly.

#### 4. EXPERIMENTATION

Extensive research has been done in order to develop our model which uses a hybrid CNN-LSTM network to generate descriptions for images.

##### 4.1 DATASET

This model has been trained on the Flickr 8k dataset. This dataset contains 8000 images each with 5 captions for each image all of them describing the image in different ways. Having different captions also allows the model to generalize better. These images are classified into the following :

Training Set	6000 images
Validation Set	1000 images
Test Set	1000 images

Table 1: Plots the average caption lengths from the training images.

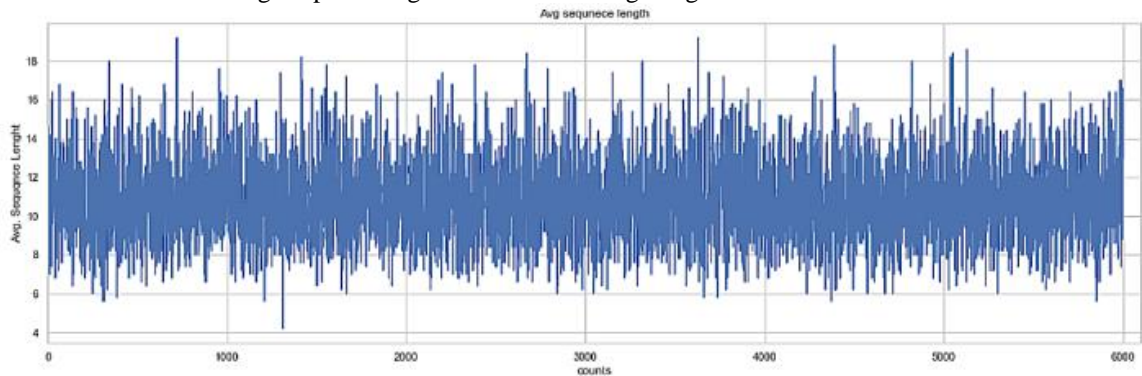


Figure 5: Plot of the average caption lengths

##### 4.2 PREPROCESSING AND PRELIMINARY ANALYSIS

We start preprocessing the dataset by removing words that occur less than 10 times since they don't carry much necessary information. Figure 6 plots the distribution of the word counts and Figure 7 plots the top most frequently occurring words.

For each image in the dataset there are 5 captions. To simplify the prediction we encode the captions into a sequence as “starting sequence” + “caption” + “ending sequence” and map them to their respective images as shown in Figure 10.

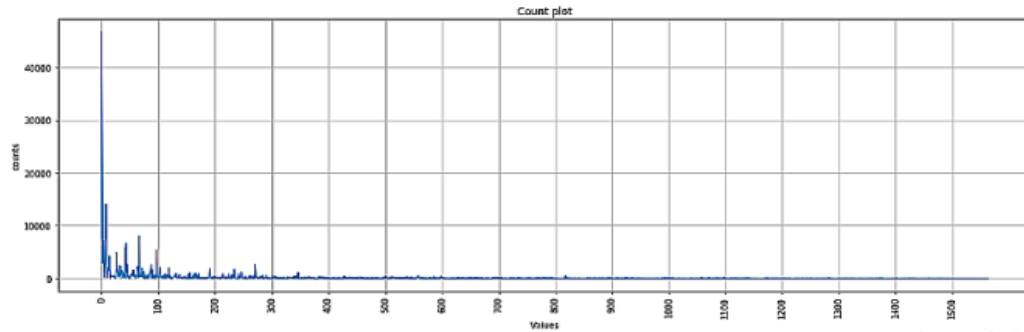


Figure 6: Distribution of word count.

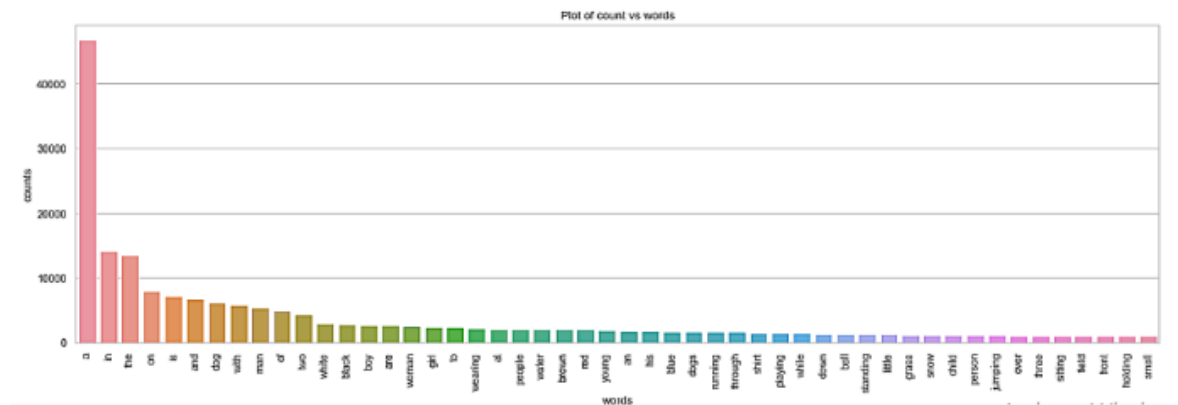


Figure 7 : Count of the most frequent words

### 4.3 IMPLEMENTATION

Every image is converted into a fixed-sized vector which can then be fed as input to the neural network. Our objective here is to get a fixed-length informative vector for each image. To extract feature vectors of the images, we opt for transfer learning by using the InceptionV3 model created by Google Research [13]. This outputs a 2048 length feature vector for every image as demonstrated in Figure 8.

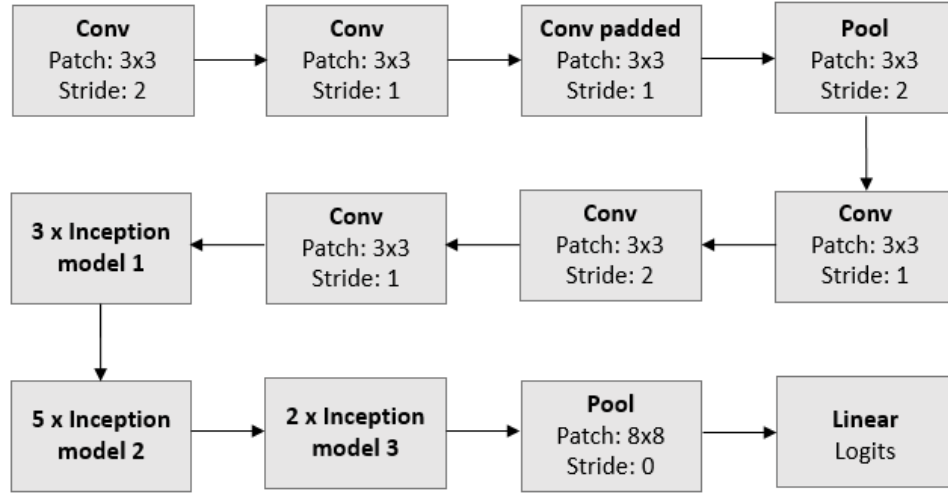


Figure 8 : Feature extraction using the InceptionV3 model.

A vocabulary of captions is made for the training images and we frame caption prediction as a supervised learning problem where we have a set of data points  $D = \{X_i, Y_i\}$ , where  $X_i$  is the feature vector of data point 'i' and  $Y_i$  is the corresponding target variable. In this way, the data matrix of one image can be summarized as in Table 2.

	X <sub>i</sub>		Y <sub>i</sub>
i	Image feature Vector	Partial caption	Target Word
1	image_1	startseq	the
2	image_1	startseq the	black
3	image_1	startseq the black	cat
4	image_1	startseq the black cat	sat
5	image_1	startseq the black cat sat	on
6	image_1	startseq the black cat sat on	grass
7	image_1	startseq the black cat sat on grass	endseq

Table 2 : A sample data matrix

It can be inferred that it is not just the image feature vector but also the partial captions of each data point that are fed to the model which helps in predicting the next word in the described sequence. An LSTM network will be reading these partial captions. Data generators are used to create new data as it is needed. A Keras generator is used so that we do not have to store the entire dataset in the memory at once and data can be generated in batches. So, in every iteration, we calculate the loss on a batch of data points to update the gradients.

Every word (index) will be mapped to a 200-long vector and for this purpose, we will use a pre-trained GLOVE Model [14]. For all the 1652 unique words in the vocabulary, we create an embedding matrix which will be loaded into the model before training. Figure 9 shows a summary of our proposed model.

```
: caption_model.summary()
```

Model: "model\_1"

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 34)]	0	
input_2 (InputLayer)	[(None, 2048)]	0	
embedding (Embedding)	(None, 34, 200)	330400	input_3[0][0]
dropout (Dropout)	(None, 2048)	0	input_2[0][0]
dropout_1 (Dropout)	(None, 34, 200)	0	embedding[0][0]
dense (Dense)	(None, 256)	524544	dropout[0][0]
lstm (LSTM)	(None, 256)	467968	dropout_1[0][0]
add (Add)	(None, 256)	0	dense[0][0] lstm[0][0]
dense_1 (Dense)	(None, 256)	65792	add[0][0]
dense_2 (Dense)	(None, 1652)	424564	dense_1[0][0]

Total params: 1,813,268  
 Trainable params: 1,813,268  
 Non-trainable params: 0

Figure 9: Model Summary

As the model is trained the weights will be updated using the backpropagation algorithm and the model will learn to output a word, given an image feature vector and a partial caption. So in summary :

Input\_1 -> Partial Caption

Input\_2 -> Image feature vector

Output -> An appropriate word with the highest probability, next in the sequence of partial captions provided in the input\_1

For a given feature vector of the image and partial caption as input to the model, we get a distribution of probability over all the words in the vocabulary. The word corresponding to the index of maximum probability is the predicted word and the sequence builds up in a similar way.

We stop predicting when “endseq” appears which denoted end sequence

Figure 10 shows a sample of the captions generated from our CNN-LSTM model for unseen images.



Figure 10 : Image descriptions generated by the model.

For the generated captions the corresponding speech has to be synthesized. For this purpose a deep learning architecture similar to Tacotron 2 as pointed out in [24]. Mel Spectrogram predictions are made over the generated captions.



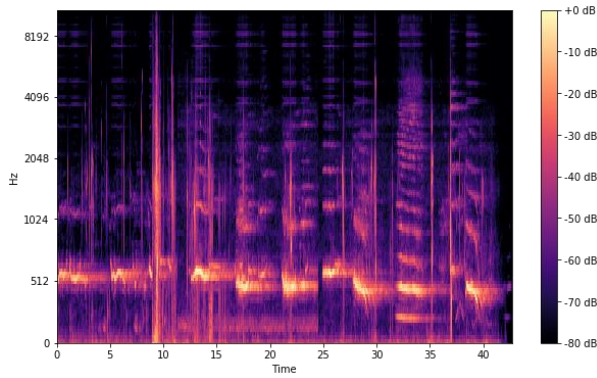


Figure 11 : Mel Spectrogram prediction from captions

We use a modified version of the WaveNet architecture from [25] to invert the mel spectrogram feature representation into time-domain waveform samples. Instead of predicting discretized buckets with a softmax layer, we follow PixelCNN++ [26] and Parallel WaveNet [27] and use a 10-component mixture of logistic distributions (MoL) to generate 16-bit samples at 24 kHz. This is the final result that speaks out the summary of the input images.

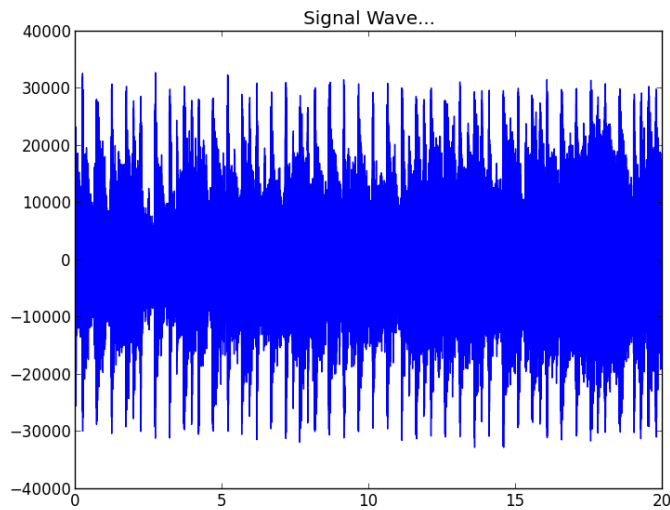


Figure 12 : Waveforms generated by WaveNet vocoder



Figure 13 : Speech generated for a single caption

## 5. CONCLUSION

The results achieved with the implementation of the hybrid algorithm can reach impressive heights, considering the training has been done based on a trained dataset, using purely supervised learning. We hope our algorithm can be a starting point for more improvements in this area, and modified to fit the daily human needs. What initially started as a method to caption images with a simple hybrid network grew bigger by adding a text to speech synthesis, mainly to target the needs of physically handicapped people, especially




those with visual impairments. It can also be used in corporate scenarios where a computer can take care of drawing conclusions from a bar chart of data. Furthermore, text to speech synthesis can be mutated to give an ASL output on a visual output actuator to help hearing-impaired people. There are endless possibilities in this spectrum and the model can be deepened, adding multiple levels to the network, to reach the desired error rate (presumably under 10%).

## REFERENCES

- [1]Soh, M. (2016). Learning CNN-LSTM architectures for image caption generation. *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep.*
- [2] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).
- [3] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010, September). Every picture tells a story: Generating sentences from images. In *European conference on computer vision* (pp. 15-29). Springer, Berlin, Heidelberg.
- [4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.Chicago
- [5]Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.
- [6]Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q. V., and Ng, A. Y. Large scale distributed deep networks. In *NIPS*, pp. 1232–1240, 2012.
- [7]Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *NIPS*, pp. 1106–1114, 2012
- [8] Singh, A., & Sharma, D. K. (2020). Image Collection Summarization: Past, Present and Future. In *Data Visualization and Knowledge Engineering* (pp. 49-78). Springer, Cham.
- [9] Wurtz R. H. (2009). Recounting the impact of Hubel and Wiesel. *The Journal of physiology*, 587(Pt 12), 2817–2823.
- [10]Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011, June). Flexible, high performance convolutional neural networks for image classification. In *Twenty-Second International Joint Conference on Artificial Intelligence*.
- [11]Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
- [12]Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10), 2222-2232.
- [13]Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).
- [14]Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
- [15]Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4565-4574).
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2
- [17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

- [18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR, 2014. 1, 2
- [19] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. arXiv preprint arXiv:1412.1441, 2014. 1, 2
- [20] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. ICLR, 2014. 1, 2
- [21] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Baby talk: Understanding and generating simple image descriptions,” in CVPR, 2011.
- [22] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” in ECCV, 2010.
- [23] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, “Collective generation of natural image descriptions,” in ACL, 2012.
- [24] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Saurous, R. A. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779-4783). IEEE.
- [25] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, “A sequence-to-sequence model for raw audio,” CoRR, vol. abs/1609.03499, 2016.
- [26] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, “PixelCNN++: Improving the PixelCNN with modifications,” in Proc. ICLR, 2017. [28] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Hassabis, “Parallel WaveNet: Fast HighFidelity Speech Synthesis,” CoRR, vol. abs/1711.10433, Nov. 2017.
- [27] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Dries, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walter, “Fast HighFidelity Speech Synthesis,” CoRR, vol. abs/1711.10433, Nov. 2017.

## BIOGRAPHIES OF AUTHORS

	Roshan varghese is a student of Computer Science Engineering in Christ University, 2017-2021.
	Regulagadda Durga Srilekha is a student of Computer Science Engineering in Christ University, 2017-2021.
	S.Yoshita Manavi is a student of Computer Science Engineering in Christ University, 2017-2021.