# END TO END DEEP NETWORK FOR IMAGE TO AUDIO CONVERSION

**Claire Vania, Nithin GK, Rohit Kumar, Vibashan VS**
cvania1@jhu.edu, ngopala2@jhu.edu,rkumar44@jhu.edu, vvishnu2@jhu.edu

## 1 ABSTRACT

Image to speech synthesis plays a crucial role in assisting blind people. Image to speech synthesis aims to synthesize intelligible and natural speech given an input image. In this work, we propose an end-to-end encoder-decoder-based architecture for the image to speech synthesis where we attempt to generate audio from image directly without generating intermediate text first. Further, we employ a transformer network to understand long-range dependencies in images for synthesizing better speech. We will analyze our model performance by evaluating it on a standard benchmark dataset and will compare with existing state-of-the-art methods.

## 2 LITERATURE REVIEW

Automatically generating captions of an image is a task very close to the heart of scene understanding — one of the primary goals of computer vision. Not only must caption generation models be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. For this reason, caption generation has long been viewed as a difficult problem. It is a very important challenge for machine learning algorithms, as it amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language. Despite the challenging nature of this task, there has been a recent surge of research interest in attacking the image caption generation problem. This section provides a brief overview of the existing work carried out in the field of image to text captioning.

The early attempts [Mao et al. (2014),Vinyals et al. (2015) ] uses the encoder-decoder paradigm that firstly utilizes CNN to encoder image and then adopts RNN based decoder to generate the output word sequence. After that, a series of innovations have been proposed to boost image captioning by encouraging more interactions between the two different modalities via attention mechanism [Cho et al. (2015)]. In particular, [Xu et al. (2015)] integrates soft and hard attention mechanism into LSTM based decoder, aiming to select the most relevant image regions for word prediction at each decoding stage. [You et al. (2016)] presents an algorithm that learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of recurrent neural networks. The selection and fusion form a feedback connecting the top-down and bottom-up computation. Instead of fully performing visual attention as in [Xu et al. (2015)], [Lu et al. (2017)] proposes an adaptive attention model that dynamically decides whether to attend to image regions at each decoding stage. Furthermore, bottom-up and top down attention mechanism [Anderson et al. (2018)] have been used extensively in image captioning and visual question answering (VQA) to enable deeper image understanding through fine-grained analysis and even multiple steps of reasoning. In this work, a combined bottom-up and top-down attention mechanism has been proposed that enables attention to be calculated at the level of objects and other salient image regions. [Qin et al. (2019)] presents Look Back (LB) method to embed visual information from the past and Predict Forward (PF) approach to look into future. LB method introduces attention value from the previous time step into the current attention generation to suit visual coherence of human. PF model predicts the next two words in one time step and jointly employs their probabilities for inference. Then the two approaches are combined together as LBPF to further integrate visual information from the past and linguistic information in the future to improve image captioning performance. Later on, the most recently proposed attention on attention module [He et al. (2016)] enhances visual attention by further measuring the relevance between the attention result and the query.

As mentioned in our project proposal, we aim to convert source modality i.e image to target modality i.e speech, using text as bridge between them. Our aim was to convert images to text using some state of the art method and from text we can synthesise speech using some state of art like tacotron2

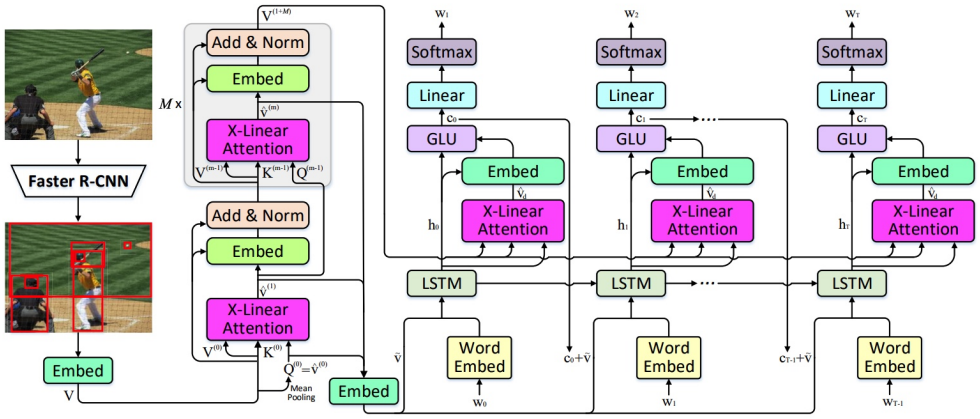


Figure 1: Overview of our X-Linear Attention Networks (X-LAN) for image captioning. Faster R-CNN is firstly utilized to detect a set of image regions. Next, a stack of X-Linear attention blocks are leveraged in image encoder to encode the region-level features with the higher order intra-modal interaction in between, leading to a set of enhanced region-level and image-level features.

Shen et al. (2018), ultimately using the best of the two worlds. This same approach was done by Ma et al. (2019) by learning multimodal representations. In their proposed approach, the author trains a multimodal information bottleneck with disjointed image to text and text to speech datasets. And during inference(speech synthesis process) they use skip modal generation so that the shared modality that is text is skipped during the inference process. Now, the problem with the proposed approach was that, for training, the model still requires text. Additionally, as mentioned in the paper Hsu et al. (2020), the The resulting captions were not eval- uated for fluency, naturalness, or intelligibility, and the BLEU scores in terms of the unsupervised units were very low (0.014 on the MSCOCO test set).

In the same year Hasegawa-Johnson et al. (2017) proposed a more direct and joint(end to end approach) approach to convert image to speech,using text as the bridge. To begin with, the paper uses MSCOCO datasetHavard et al. (2017), and for training firstly they train XNMT image to speech model, by taking image embeddings(generated by VGGish network) and L1 and L2 phones (generated by Kaldi Povey et al. (2011)). Then train a clustergen TTS(text to speech) model Cahyaningtyas & Arifianto (2016) to convert those units generated by XNMT model to convert to synthesise speech. While generating the audio, image features are converted to speech unit sequence. The speech unit sequence is then used by TTS as input to generate the speech utterances. However, their approach uses L1 and L2 phones which still need textual information to generate phones from the speech.

Additionally this year, paper by Wang et al. (2021) proposed a text free end to end image to speech model, which simplifies the task by using a pair of image and synthesized speech generated from a single speaker TTS model to reduce the acoustic variation. For this project, however, we will mainly try to implement and improve upon the work by Hsu et al. (2020) where they directly convert image to speech, by first converting image to some discrete units and then those discrete units are being converted into speech. The main contributions of this paper in the field of image to speech captioning was, first it was the first fluent image to speech synthesis that does not rely on text. Second, this project collected/curated MSCOCO dataset Havard et al. (2017)[1], which contains 600,000 spoken audio captions. This paper will be discussed in detail in the proposed methodology section.

[1]for data: https://github.com/William-N-Havard/SpeechCoco

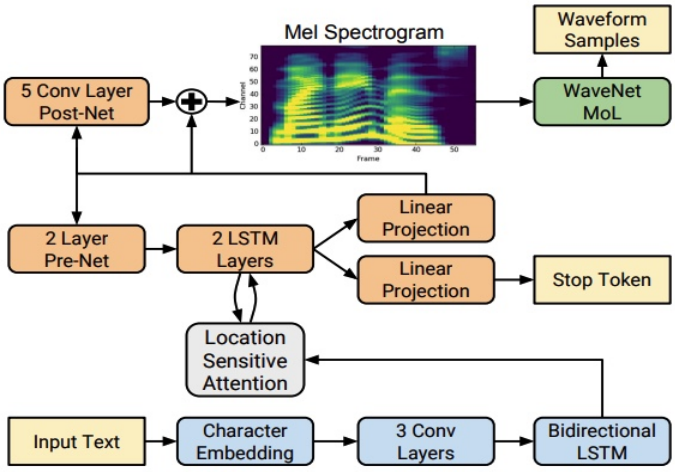


Figure 2: Block diagram of the Tacotron 2 system architecture

# 3 Proposed Methodology

## 3.1 Distributed Approach

### 3.1.1 Image to Text

We create a unified X-Linear attention block for image captioning, which models the 2nd order interactions with both spatial and channel-wise bilinear attention. The higher and even infinity order feature interactions can be readily modeled via staking multiple X-Linear attention blocks and equipping the block with Exponential Linear Unit (ELU). Dataset to be used is COCO [Lin et al. (2014)]. The whole COCO dataset contains 123,287 images, which includes 82,783 training images, 40,504 validation images, and 40,775 testing images. Each image is equipped with five human-annotated sentences. Bilinear pooling is an operation to calculate outer product between two feature vectors. Such technique can enable the 2nd order interaction across all elements in feature vectors and thus provide more discriminative representations than linear pooling. Considering the successes of bilinear pooling applied in fine-grained visual recognition [Gao et al. (2016), Yu et al. (2018)] or visual question answering [Fukui et al. (2016), Kim et al. (2018)], we fully capitalize on bilinear pooling techniques to construct a unified attention module (X-Linear attention block) for image captioning. Such design of X-Linear attention block strengthens the representative capacity of the output attended feature by exploiting higher order interactions between the input single-modal or multi-modal features. As mentioned, X-Linear attention is a unified attention block it is feasible to plug X-Linear attention blocks into image encoder and sentence decoder to capture higher order intra- and inter-modal interactions for image captioning. The sentence decoder aims to generate the output sentence conditioned on the enhanced image-level and region-level visual features induced by the image encoder. To further encourage high order inter-modal interactions between visual content and natural sentence, we integrate our X-Linear attention block into attention-based LSTM decoder to perform multi-modal reasoning.

### 3.1.2 Text to Speech

The task of generating natural sounding speech from text remains a challenging problem to be solved. Deep learning based TTS systems are the current state-of-the-art in terms of producing natural sounding speech. Among the various deep learning based end-to-end models, in our project we used Tacotron2 Shen et al. (2018) and WaveGlow Prenger et al. (2019). Tacotron2 model is an encoder-attention-decoder setup where 'location sensitive attention' is used. The first part is an encoder which converts the character sequence into word embedding vector. This representation is later consumed by the decoder to predict spectrograms. To generate timedomain waveforms from

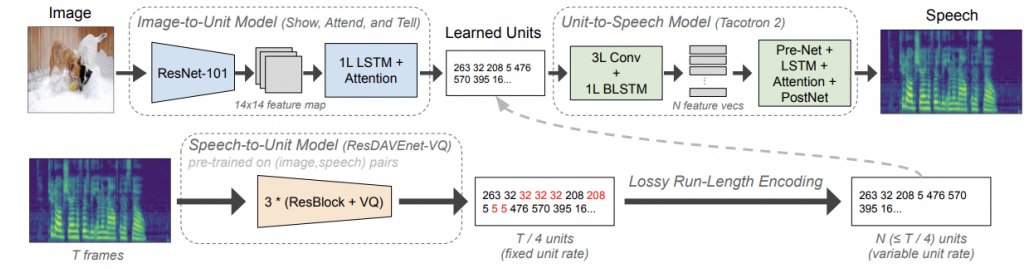


Figure 3: A model similar to our end objective.

the spectograms predicted by the Tacotron2 model, the WaveGlow Prenger et al. (2019) vocoder is used.

For our project we used ESPnet toolkit Watanabe et al. (2018) with the Pytorch backend Paszke et al. (2017). The dataset used to train the model is LJspeech dataset Ito & Johnson (2017).

### 3.2 End to End Approach

For the implementation of end to end image to speech conversion, we will be implementing and improving the work by Hsu et al. (2020), which is shown in the figure 3. The entire model can be divided into three subsections, that is I2U(image to unit),S2U(speech to unit) and U2S(unit to speech). To begin with, given the image, it is being passed through the image to unit model in order to predict the unit that is $P(U|I)$. In the paper, open implementation of Show, Attend and Tell Xu et al. (2015) is used with soft attention, where rather then CNN encoder, ResNet model(pre trained on imageNet for image classification) is being used.

Now for the S2U(speech 2 unit), ResDAVEnet-VQ "2 $\rightarrow$ 2, 3" model and the WaveNet-VQ (PA) model reported by Harwath et al. (2019) is being used. Both models learn discrete representations for speech and are used to transcribe speech into a sequence of units in this paper. And finally, given the linguistic symbol sequence U, units are being passed through the tacotran 2 Shen et al. (2018) model , which is being trained on these linguistic units and used to synthesise speech that is $P(S|U)$. The main advantage of this approach was that in the entire image to speech, they do not generate text. For our project, we wanted to use the model architecture somewhat like the given model, but in the end, we wanted to train the model E2E i.e trying to predict $P(S|U)$, in one shot.

#### 3.2.1 Dataset

In the original paper Hsu et al. (2020) 5 different dataset are being used. But for our implementation, we are going to use 3 datasets, that is using Spoken COCO Havard et al. (2017) for the I2U unit, LJspeech Ito & Johnson (2017) for U2S and Paces Audio Harwath & Glass (2017) for S2U. Different datasets are being used to train different models, to check the robustness of the model.

## 4 Results of Phase 1: Implementation of Distributed Approach

#### 4.0.1 Image to Text

Table 1 summaries the performance comparisons between the state-of-the-art models and our proposed X-LAN on the offline COCO Karpathy test split. The implementation can be found here[2]

[2] https://github.com/JDAI-CV/image-captioning

| | Cross Entropy Loss | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | B@1 | B@2 | B@3 | B@4 | M | R | C | S |
| **LSTM** | - | - | - | 29.6 | 25.2 | 52.6 | 94.0 | - |
| **X-LAN** | 78 | 62.3 | 48.9 | 38.2 | 28.8 | 58 | 122 | 21.9 |

Table 1: Performance comparisons on COCO Karpathy test split, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr and SPICE scores. All values are reported as percentage (%).

#### 4.0.2 TEXT TO SPEECH

We use the ESPnet toolkit Watanabe et al. (2018) for converting captions(text) to speech(using tacotron 2), trained on LJspeech dataset. Some of the audio, synthesised using this this model can be found here.[3]

#### 4.0.3 IMAGE TO SPEECH CAPTIONS

After using both the distributed networks, that using Image to text model to generate captions(text) and using text to speech model to use the captions to synthesise the audio, which can be found here.[4]

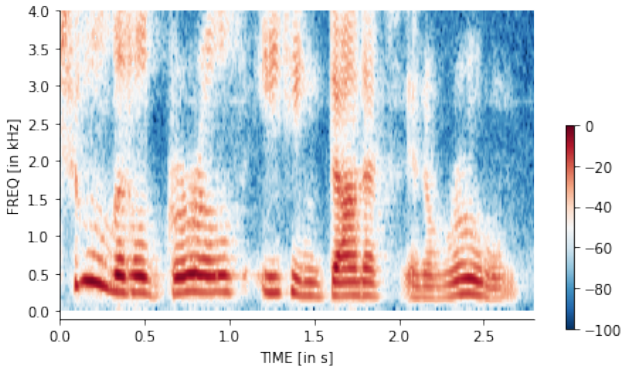


Figure 4: (a)Caption generated by X-LAN: : two little girls eating donuts in a room
(b) Spectrogram of speech synthesised for the caption

## 5 PHASE 2: IMPLEMENTING E2E IMAGE TO SPEECH [5]

- To implement and improve upon the work of Hsu et al. (2020).
- Even though Hsu et al. (2020), directly convert image to speech without explicitly using text as an intermediate representation, the training process for discrete speech units still needs a full amount of parallel image-speech data, which in most of the scenario difficult to find. So our final goal will be, if we remove/replace text from the training phase and by some means train the model without the need for parrallel text.

[3] https://drive.google.com/drive/folders/1MZQnDig9KVMqmttCMbpFJSvHZYgG89VQ?usp=sharing

[4] Image to Speech caption results: https://drive.google.com/drive/folders/1g-m_z27m5RQYYwRoTW7XB6keZge-4TbY?usp=sharing

[5] will implement in end sem

## REFERENCES


Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

Elok Cahyaningtyas and Dhany Arifianto. Synthesized speech quality of indonesian natural text-to-speech by using hts and clustergen. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 110–115. IEEE, 2016.

Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 317–326, 2016.

David Harwath and James R Glass. Learning word-like units from joint audio-visual analysis. *arXiv preprint arXiv:1701.07481*, 2017.

David Harwath, Wei-Ning Hsu, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. *arXiv preprint arXiv:1911.09602*, 2019.

Mark Hasegawa-Johnson, Alan Black, Lucas Ondel, Odette Scharenborg, and Francesco Ciannella. Image2speech: Automatically generating audio descriptions of images. *Proceedings of ICNLSSP, Casablanca, Morocco*, 2017.

William Havard, Laurent Besacier, and Olivier Rosec. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. *arXiv preprint arXiv:1707.08435*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. *arXiv preprint arXiv:2012.15454*, 2020.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383, 2017.

Shuang Ma, Daniel McDuff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7598–7607, 2019.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

A. Paszke, S. Gross, et al. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621. IEEE, 2019.

Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. Look back and predict forward in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8367–8375, 2019.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 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, 2018.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Xinsheng Wang, Siyuan Feng, Jihua Zhu, Mark Hasegawa-Johnson, and Odette Scharenborg. Show and speak: Directly synthesize spoken description of images. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4190–4194. IEEE, 2021.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.

Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 574–589, 2018.