A Deep Learning Approach for Nepali Image Captioning and Speech Generation

Sagar Sharma, Samikshya Chapagain, Sachin Acharya, Sanjeeb Prasad Panday*
Department of Electronics and Computer Engineering-Pulchowk Campus-Institute of Engineering,
Tribhuvan University, Lalitpur, Nepal

***Abstract*—This article introduces a novel approach for Image-to-Speech generation that aims in converting images into textual captions along with spoken descriptions in Nepali Language using deep learning techniques. By leveraging computer vision and natural language processing, the system analyzes images, extracts features, generates human-readable captions, and produces intelligible speech output. The experimentation utilizes state-of-the-art transformer architecture for image caption generation complemented by ResNet and EfficientNet as feature extractors. BLEU score is used as an evaluation metric for generated captions. The BLEU scores obtained for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 n-grams are 0.4852, 0.2952, 0.181, and 0.113, respectively. Pretrained HifiGaN(vocoder) and Tacotorn2 are used for text to speech synthesis. The proposed approach contributes to the underexplored domain of Nepali-language AI applications, aiming to improve accessibility and technological inclusivity for the Nepali-speaking population.**

***Keywords*—Image captioning; speech generation; image-to-speech generation; deep learning; BLEU score; HiFiGaN; TTS**

I. Introduction

Image caption generation is a crucial task within the realms of Computer Vision and Natural Language Processing, entailing the creation of textual descriptions corresponding to images. This process serves a variety of purposes, from aiding visually impaired individuals to indexing images and facilitating image-based search engines. It is a pivotal aspect of scene understanding in Computer Vision, demanding not only the identification of objects within an image but also their coherent expression in natural language. While traditional object detection and classification tasks exist, caption generation poses a more intricate challenge, necessitating the fusion of Computer Vision techniques for content comprehension and feature extraction with Natural Language Processing methods for sequential description generation.

Recent advancements have demonstrated the effectiveness of transformer-based models over conventional CNN-RNN architectures for image captioning. Research in languages such as Hindi and Bengali has achieved notable success using transformer networks. However, Nepali remains an underexplored language in this domain due to the scarcity of labeled datasets. This study bridges this gap by employing a deep learning-based image-to-speech generation approach, integrating image captioning with speech synthesis to provide a holistic multimodal AI system.

The paper is organized as follows: Section II reviews related works, discussing existing methodologies and their limitations. Section III introduces the proposed approach, detailing model architecture and dataset preparation. Section IV explains the methodology in detail, followed by Section V, which describes the experimental setup. Section VI presents the results, while Section VII provides a discussion. Finally, Section VIII concludes with key insights and future research directions.

*Corresponding authors.

II. Related Works

The field of image captioning has witnessed significant progress through various related works. One notable contribution is the "Show and Tell" [1] model, which introduced a framework for image captioning using a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This approach first encodes the image using a CNN to extract visual features and then generates a caption using an RNN. It laid the foundation for subsequent advancements in image captioning by demonstrating the feasibility of combining deep-learning models to generate captions for images.

Several studies have delved into image captioning across languages, highlighting the superiority of transformer-based models over conventional CNN-RNN architectures. Mishra et al. (2021)[2] presented a transformer-based encoder-decoder architecture for Hindi image captioning, emphasizing the drawbacks of RNN models. They obtained substantial BLEU scores, marking the efficacy of the transformer in generating accurate captions. Similarly, Ami et al. (2020) [3] explored Bengali image captioning using CNN and transformer networks, demonstrating improved performance compared to earlier models. These studies utilized pre-trained CNN models like ResNet-101, InceptionV3, and Xception, along with various datasets such as BanglaLekha and Flickr8k, showcasing enhanced BLEU and METEOR scores.

Image caption generation in various languages has seen significant progress, particularly in languages like English, Hindi, and Bengali. However, Nepali, due to limited datasets and research focus, remains an understudied language in this domain. Adhikari and Ghimire (2019)[4] introduced Nepali image captioning using encoder-decoder models, but with limited performance owing to the use of traditional CNN-RNN architectures. Addressing this gap, this study by Subedi and Bal (2022) [5] pioneers Nepali image captioning utilizing a CNN-Transformer model, leveraging the strengths of transformer networks and their effectiveness in handling NLP tasks.

Tacotron [6], which combined an encoder-decoder framework with a sequence-to-sequence model to generate high-quality speech. Another influential contribution is WaveNet [7]

which introduced a deep generative model capable of generating high-fidelity audio waveforms. Building upon Tacotron developed Tacotron 2 [6] in 2018, incorporating a WaveNet vocoder to enhance the naturalness and quality of synthesized speech. Deep Voice, a series of models introduced by Arik et al. [8] in 2017, employed deep convolutional and recurrent neural networks for speech synthesis. Transformer TTS [9] proposed in 2019, harnessed the Transformer architecture originally used for machine translation and showcased excellent results in generating natural and expressive speech.

In the landscape of image description, especially in the context of Nepali speech, a noticeable dearth of research exists. While textual image captioning has received some attention in the Nepali language, the adaptation of these methodologies to cater to speech-based image descriptions in Nepali remains largely unexplored. This gap in the literature signifies an opportunity for innovative research and development. In light of this research void, the present project assumes significance by actively addressing the under representation of Nepali speech-based image description.

The absence of sufficient research and publications in the domain of Nepali speech-based image description underscores the critical need for initiatives. By recognizing and addressing this gap in the literature, this study stands as a pivotal endeavor to advance image understanding technologies tailored to Nepali speech, promising to improve accessibility and technological inclusivity for the Nepali-speaking population.

III. Proposed Work

This article emphasizes not only in the image caption generation but extends the model for generating the fluent and natural speech for the generated textual captions in Nepali Language. For understanding the visual content of the images, various state of the art architectures of Convolutional Neural Network are used for the image captioning tasks. Architectures such as VGGNet, EfficientNet, ResNet, InceptionNet have produced remarkable results which extracts various features from the image and thus generates a feature map that captures the important representations of the image.

Further for the caption generation various RNN architectures are used because of its robustness in sequential language tasks. The use of Long Short-Term Memory and Gated Recurrent Units have been overshadowed by the transformer architecture because of the attention mechanism that enables to selectively attend the different parts of the input data, assigning varying degrees of importance to different elements.

The Pretained Text to Speech [10] serves as a crucial element in the model pipeline, providing synthesized speech outputs based on textual inputs. Specifically, Tacotron2, a state-of-the-art Text to Speech (TTS) model is employed, which has been pre-trained on a diverse dataset encompassing a wide range of linguistic features and speech patterns. By integrating a pre-trained TTS system into the framework, the ability to generate high-quality spectrograms from input text is caplitalized, thereby streamlining the process of speech synthesis.

Moreover, the utilization of a pre-trained TTS system offers several advantages, including reduced training time and resource requirements, as well as enhanced synthesis quality owing to the model's extensive training on large-scale datasets. By incorporating Tacotron2 into the research, the image-to-speech generation pipeline is benefited from cutting-edge TTS technology, thereby enabling to achieve superior performance in generating natural and intelligible speech outputs for Nepali language inputs [11] to [16].

IV. Methodology

The methodology involves a combination of image processing, feature extraction, and natural language processing. Image processing and feature extraction includes the utilization of architecture of ResNet and EfficientNet. The extracted features are further used in the transformer architecture for predicting captions in Nepali Language. BLEU score metric has been used as an evaluation metric for n-grams prediction up to 4-grams. Pretrained text to speech model has been utilized to convert the textual captions to Nepali speech.

A. Dataset Preparation

Flickr8k dataset is used in the research which is openly available in the Internet. The dataset consists of real life images specially human and animal activities and each image consisting of four or five English annotated captions. The English captions are further annotated using Google Translation API into Nepali captions.

B. Image Preprocessing

The images are normalized by resizing them into a consistent size and resolution. Augmentation techniques such as rotation, flipping and blurring is used in order to generalize the images for better feature extraction.

C. CNN Feature Extraction

Convolutional Neural Network architecture ResNet and EfficientNet are used for high level visual feature extraction. EfficientNetB7 and ResNet50 architecture is used which is fine-tuned for Flicker8k dataset using pretrained weight from the ImageNet dataset and the particular fine-tuned weights are utilized. The extracted features capture meaningful visual representations.

D. Caption Generation

The captions are tokenized and vectorized as required before training. The transformer architecture chosen so as to utilize its ability of attention mechanism effectively. The model thus learns to associate the visual features with Nepali captions during training. In sharp contrast to the orthodox single layered transformer implementation, this article introduces the utilization of fine tuned nepali embedding layers. Similarly, layered transformer model is also used for caption generation.

E. Evaluation

The performance of the trained model is evaluated using the BLEU (Bilingual Evaluation Understudy) metric. The metric evaluates the predicted caption by comparing it with the reference captions on the basis of the n-grams similarity, where the sequential groups of words such as unigrams (single words), bigrams (two word sequence), trigrams (three word sequence) and so on.

F. Speech Synthesis

The speech synthesis model incorporates Tacotron which is a text-to-speech (TTS) model that generates mel-spectrograms from input text. HiFi-GAN is a high-fidelity generative adversarial network designed for audio waveform generation. Tacotron provides accurate and expressive mel-spectrograms, while HiFi-GAN transforms them into high-fidelity speech waveforms.The combined approach allows for fine-grained control over synthesized speech and produces natural-sounding output.

The overall workflow of the proposed methodology is illustrated in Fig. 1.

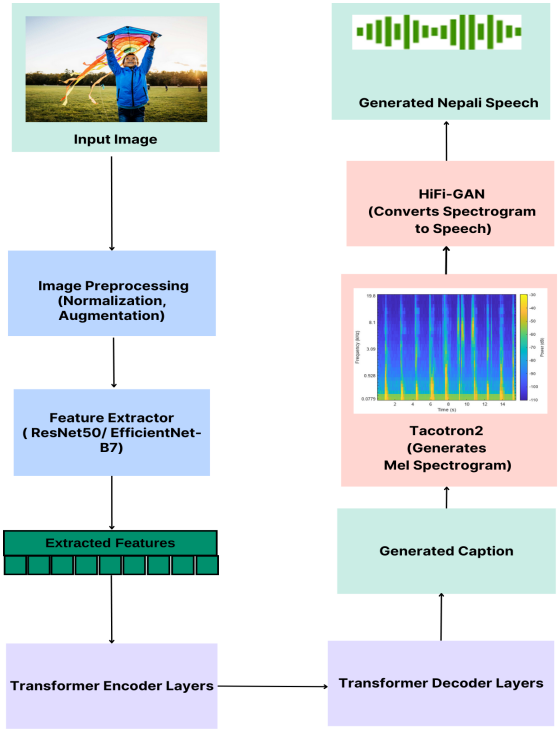

Fig. 1. Proposed methodology workflow.

V. Experimental Setup

A. Datasets

Flickr 8k dataset is utilized for training of image captioning transformer model. The dataset consist of 8k images mostly representing human, dogs and their activities. The data shuffled using random seed 42 produced best accuracy. The original dataset consists of four or five English captions for each image. The Google Language Translation API is used for the corresponding conversion into Nepali captions. Poorly translated captions is removed by manual examinations. The dataset is divided into training and validation set as shown in Table I below:

TABLE I. Split of Dataset

Dataset	Training Split	Validation Split
Flickr8k	6.4k	1.6k

B. Model Architectures

Model uses transformer based encoder decoder architecture for image caption generation. Model utilizes one of the two CNN architectures for feature extraction: ResNet50 and EfficientNetB7. The output of feature extractor is passed to the transformer encoder for further processing differentiating from the previous works which was primarily based on the single layered transformer model with the consideration of decoder part only. The layered transformer and learned embedding models are utilized as two new models for image caption generation. Both layered transformers and learned embedding uses EfficientNetB7 as the feature extractor. EfficientNetB7 uses less computation power and produces similar accuracy. For the complex and computationally expensive models like layered transformer and learned embeddings, EfficientNetB7 was implemented to reduce the overall computation requirements. Fig. 2 illustrates the architecture of single-layered transformer.

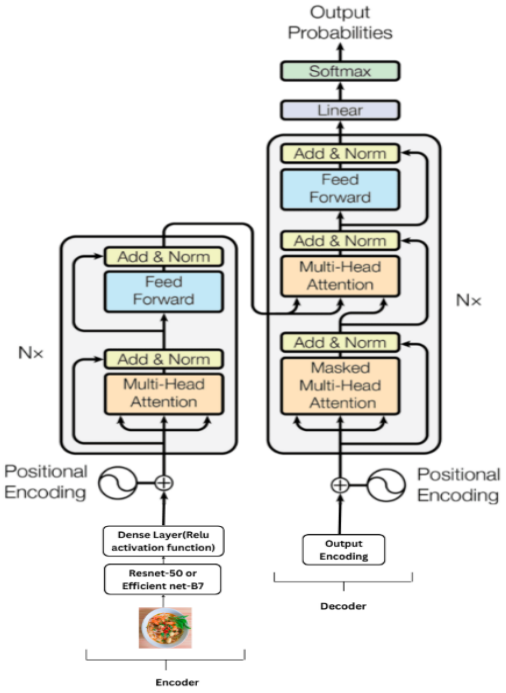

Fig. 2. Single layer transformer architecture.

1) Feature extractor: CNN architectures Resnet50 and EfficientNetB7 are utilized for extracting spatial features from images. Pretrained Imagenet weights are utilized for the extraction of image features. The augmentation techniques such as Random Flip, Random Rotation and Random Zoom are implemented for increasing the diversity of the training images. Classification layer is removed and output of second last layer is utilized as an input for the transformer encoder layer.

2) Transformer encoder: In order to take advantage of attention mechanism in the encoded image, encoder layer of transformer is utilized. The output of the feature extractor is passed to the positional embedding layer. The positional encoded features are normalized using layer normalization.

The output is then passed to the Dense layer with ReLU activation function. The extracted embeddings are passed to the multihead self attention layer. Introduction of dense layer with non linear activation function (ReLU) enables the encoder to learn complex images features and attention layer learns multiple embedding dimensions. The self attention in the encoded image features allows model to attend the important features of the image for generation of distinct captions.

3) Decoder: The transformer decoder layer is utilized for decoding the image caption. It generates caption using one word at each timestep. The previously generated words are passed as an input to the decoder. Vectorizer vectorizes each word. Vectorized word is passed to the positional encoding layer of the transformer. Masking is used to prevent the transformer to attend the future output. Masked output is then passed to the multihead self attention layer which highlights only the important part of input data while diminising the effect of the rest. The attended decoder input and the input from encoder are passed through the attention layer called encoder decoder attention. This layer builds the relationship between extracted features from encoder and the attention output from decoder. This relationship along with residual connection is finally used for predicting the Nepali caption. The whole architecture is trained using backpropagation techniques.

C. Compared Methods

The single layered transformer model with ResNet-50 as feature extractor is compared with pretrained embedding model and layered transformer model.

1) Pretrained embedding model: The pretrained embedding model utilizes the pretrained nepali word2vec text embedding. This pre-trained Word2Vec model had 300-dimensional vectors for more than 0.5 million Nepali words and phrases. About 20% of the embeddings from our dataset was missing in the pretrained embedding. Fine tuning of the word2vec model is performed using window size as 5. The model was trained for 10 epochs and fine tuned embedding were than utilized in the transformer model for image caption generation.

Every word is vectorized using text vectorizer. The vectorized word is mapped to 300 dimension word embedding. The embedding layer is trained along with other parameters of transformer decoder. Word embedding were utilized for finding relation between words in the training caption using already trained embedding from large corpus.

2) Layered transformer model: Single layer transformer is stacked in both encoder and decoder side to obtain multilayered transformer model. Three encoder layers and three decoder layers were stacked to search for intricate relationship between features. The complexity of the model goes on increasing from single layered transformer model. Single layered model is least computationally expensive while layered transformer model require highest computation as compared to other two models.

D. Evaluation Metrices

The performance of the trained model is evaluated using the BLEU (Bilingual Evaluation Understudy) metric.The metric evaluates the predicted caption by comparing it with the reference captions on the basis of the n-grams similarity, where the sequential groups of words such as unigrams (single words), bigrams(two word sequence), trigrams(three word sequence) and so on.

Four BLEU scores (BLEU-1, BLEU-2, BLEU-3, and BLEU4) are typically calculated in the context of image captioning. These scores evaluate merely matching grams of a certain order, such as single words (1-gram) or word pairs (2-gram or bigram), and so forth. BLEU score ranges from 0 to 1 where a score between 0.6 to 0.7 is considered to be the best achievable result but at the same time, a score between 0.3 to 0.4 is considered an understandably good translation and a score greater than 0.4 is considered high-quality translation.

The evaluation of the image-to-caption model's performance also incorporated the utilization of the sentence BLEU score. Unlike corpus BLEU, which uses the word level tokenization, sentence BLEU is modified to find BLEU Score using character-level tokenization to evaluate the quality of each caption independently.This sentence-level evaluation metric offers a more detailed assessment of caption quality. Considering the absence of lemmatizer for the nepali word corpus, the word level tokenization was not able to predict the generalizing ability of the model. For example, dog and dogs can be lemmatized easily for English corpus but due to complicated structure of Nepali corpus, root words could not be extracted properly. Such complication were removed using character level tokenization which does not significantly penalizes the prediction of similar words other than the root word such as dog and dogs in Nepali. Notably, the sentence BLEU method provided metrics that closely aligned with the benchmark evaluations.

E. Text-to-Speech

The implementation of a pretrained Text-to-Speech (TTS) model within the image-to-speech conversion framework was informed by a scholarly paper detailing Nepali Text-to-Speech synthesis using Tacotron2 for melspectrogram generation. The research method delineated in the paper involves preprocessing and tokenization of Nepali text, which is subsequently fed into a Tacotron2 model to generate melspectrograms. This model, initially trained on an established Nepali dataset, is further refined through fine-tuning on a proprietary dataset to enhance performance and adaptability to language-specific nuances. Employing incremental learning techniques ensures continual updates to the model, enabling it to accommodate evolving linguistic contexts effectively. The melspectrograms generated by the Tacotron2 model are then processed using HiFiGAN and WaveGlow vocoders to produce synthesized speech. Post-processing methods are subsequently applied to refine the output and improve its naturalness. Notably, qualitative evaluation of the synthesized speech yielded a commendable Mean Opinion Score for naturalness, signifying the efficacy of the employed approach. By leveraging the insights gleaned from this scholarly work, the integration of the pretrained TTS model into the image-to-speech conversion system has significantly enhanced the quality and naturalness of the synthesized speech output, particularly for the Nepali language.

VI. Results

The Flickr8k dataset consists of 8k images with 4-5 captions for each images. Same dataset was trained on three different models. The dataset was split into training set and validation set. 6.4k dataset are used for training set and remaining 1.6k are used for validation. The validation dataset is used as test set for generating caption and calculation of BLEU scores after the models are successfully trained. The model parameters used in the experiments are summarized in Table II.

TABLE II. Model Parameters

Parameters	Values
Batch Size	32
Embedding Dimension	512
Feed Forward Dimension	2048
Image Size	(224, 224)
Learning Rate	10^{-4}
Loss Function	Sparse Categorical Cross Entropy
Number of Heads	3
Optimizer	Adam
Sequence Length	25
Vocab Size	15,000

The corpus BLEU score is calculated for different model architectures, with the results presented in Table III. Fig. 3 displays the same data in bar chart form for better comparison.

TABLE III. Corpus BLEU Score

Model	Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Layered Transformer	EfficientNetB7	0.381	0.176	0.0664	0.02145
Learned Embeddings	EfficientNetB7	0.475	0.269	0.124	0.050
Single Layered Transformer	ResNet-50	0.446	0.268	0.184	0.139

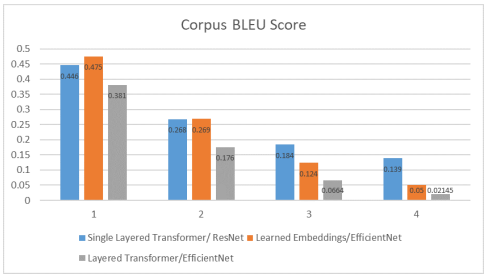

Fig. 3. Corpus BLEU score.

Similarly, the sentence-level BLEU score is calculated, providing a more granular and detailed evaluation of the generated captions. The scores are presented in Table IV, and the bar graph is shown in Fig. 4.

TABLE IV. Sentence BLEU Score

Model	Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Single Layered Transformer	ResNet-50	0.5223	0.3858	0.3013	0.2462
Learned Embeddings	EfficientNetB7	0.5217	0.3950	0.3171	0.2651
Layered Transformer	EfficientNetB7	0.4960	0.3528	0.2662	0.2125

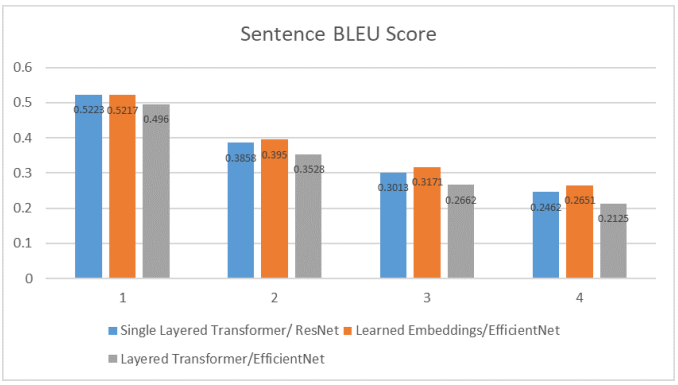

Fig. 4. Sentence BLEU score.

Fig. 5 to 8 present the generated Nepali captions alongside their English equivalents, demonstrating the Single Layered Transformer model's effectiveness in generating captions for different images.

Fig. 5. Equivalent English caption: A guy in a blue shirt is ready to kick a football.

Fig. 6. Equivalent English caption: Two dogs are running on the grass.

Fig. 7. Equivalent English caption: A big bird has landed on the water.

Fig. 8. Equivalent English caption: A dog is running on the field.

VII. Discussion

BLEU scores of layered transformer model is observed to be significantly lower than other models. This is found to be because of the overfitting issue as the increase in layer significantly increased the learnable parameters thus the feature extracted and the size of dataset used were not enough to overcome overfitting issue.

Result obtained by using learnable pretrained embedding layer worked slightly better than other models. The intricate relationship between the words are captured well using embedding layer. This is indicated by improvement in BLEU-2, BLEU-3, BLEU-4 scores. However in case of word level tokenization, BLEU-1 for learned embedding is greatest among all models but the relationship between more than two words could not be generalized well which is marked by low BLEU-3 and BLEU-4 scores.

Single layered transformer model with Resnet-50 as feature extractor is performing at similar accuracy when compared to learned embeddings. Since the complexity of model is low and the dataset is also sparse with only 8k images, training a model like single layered transformer with less learnable parameter performed remarkably well. However, Pretrained embedding could be preferable if abound dataset is available for image captioning.

The results highlight the model's ability to capture diverse scenarios with remarkable detail. In Fig. 5, it not only identifies a person but also recognizes the blue color of the shirt, highlighting its capability to capture intricate visual features. Similarly, in Fig. 6, the model accurately detects two dogs running on the grass, effectively distinguishing multiple objects and their actions. Fig. 7 showcases its ability to recognize both a large bird and the presence of water, indicating its understanding of different elements within a scene. Likewise, in Fig. 8, the model successfully identifies a dog running on a field, emphasizing its proficiency in capturing both subject and activity. These examples indicate the model's effectiveness in generating descriptive and contextually rich captions.

VIII. Conclusion

This study explored a deep learning-based approach for image captioning and speech synthesis in Nepali. Transformer-based architectures were employed for caption generation, while state-of-the-art text-to-speech models were utilized to produce spoken descriptions. Evaluation using BLEU scores demonstrated that a single-layered transformer model with ResNet-50 as the feature extractor performed competitively, while learned embeddings offered slight improvements in capturing linguistic relationships.

One of the primary challenges identified was the limited availability of high-quality Nepali image-caption datasets, which impacted model performance. Additionally, the complex morphological structure of Nepali presented difficulties in tokenization and evaluation, emphasizing the need for more refined linguistic processing techniques. Addressing these challenges in future research could lead to improvements in both caption quality and speech synthesis accuracy.

The findings contribute to the underexplored domain of Nepali-language AI applications, highlighting the potential for advancements in image-to-speech technology for low-resource languages. Future research can focus on dataset expansion, the integration of more advanced language models, and enhancements in text-to-speech synthesis to better capture the nuances of Nepali pronunciation and grammar.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," 2014, arXiv:1411.4555.

[2] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya, and A. K. Singh, "Image captioning in Hindi language using transformer networks," *Computers & Electrical Engineering*, vol. 92, p. 107114, 2021, doi:10.1016/j.compeleceng.2021.107114.

[3] A. S. Ami, M. Humaira, M. A. R. K. Jim, S. Paul, and F. M. Shah, "Bengali Image Captioning with Visual Attention," in *Proc. 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1-5, doi:10.1109/ICCIT51783.2020.9392709.

[4] A. Adhikari and S. Ghimire, "Nepali Image Captioning," in *Proc. Artificial Intelligence for Transforming Business and Society (AITB)*, 2019, vol. 1, pp. 1-6, doi:10.1109/AITB48515.2019.8947436.

[5] B. Subedi and B. K. Bal, "CNN-Transformer based Encoder-Decoder Model for Nepali Image Captioning," in *Proc. 19th International Conference on Natural Language Processing (ICON)*, New Delhi, India, 2022, pp. 86-91.

[6] Y. Wang *et al.*, "Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model," 2017, arXiv:1703.10135.

[7] J. Shen *et al.*, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," 2017, arXiv:1712.05884.

[8] S. O. Arik *et al.*, "Deep Voice: Real-time Neural Text-to-Speech," 2017, arXiv:1702.07825.

[9] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Neural Speech Synthesis with Transformer Network," 2019, arXiv:1809.08895.

[10] S. Khadka, R. G.C., P. Paudel, R. Shah, and B. Joshi, "Nepali Text-to-Speech Synthesis using Tacotron2 for Melspectrogram Generation," in *Proc. SIGUL 2023*, pp. 73-77, doi:10.21437/SIGUL.2023-16.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, arXiv:1512.03385.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," 2020, arXiv:2010.05646.

[14] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," 2014, arXiv:1406.2661.

[15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014, arXiv:1409.1556.

[16] L. Srinivasan and D. Sreekanthan, "Image Captioning-A Deep Learning Approach," 2018, [Online]. Available: https://api.semanticscholar.org/CorpusID:85556880.