IMAGE CAPTIONING

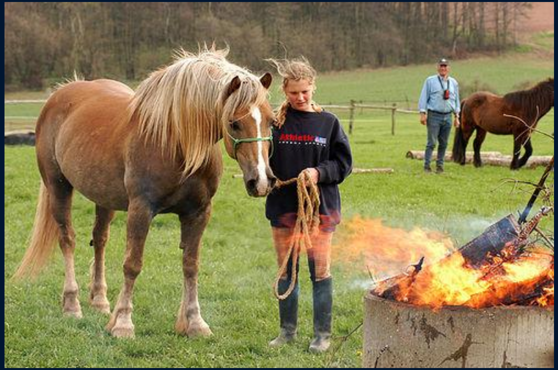
DEEP LEARNING PROJECT GROUP - U4

AADITYA GUPTA (2020552)
HARJEET SINGH YADAV (2020561)
HARSH VARDHAN SINGH (2020202)

PROBLEM STATEMENT

The problem is to generate natural language captions for a given image using deep learning models. Given an input image, the goal is to develop a model that can generate a coherent and descriptive caption that describes the object's, attributes, and relationships depicted in the image. This is a challenging task as it involves understanding the visual content of the image and generating language that accurately describes that content in a meaningful and grammatically correct way. The success of the model will be evaluated based on the quality and relevance of the generated captions.

Eg.



→

'a blonde horse and a blonde girl in a black sweatshirt are staring at a fire in a barrel ',

'a girl and her horse stand by a fire ',

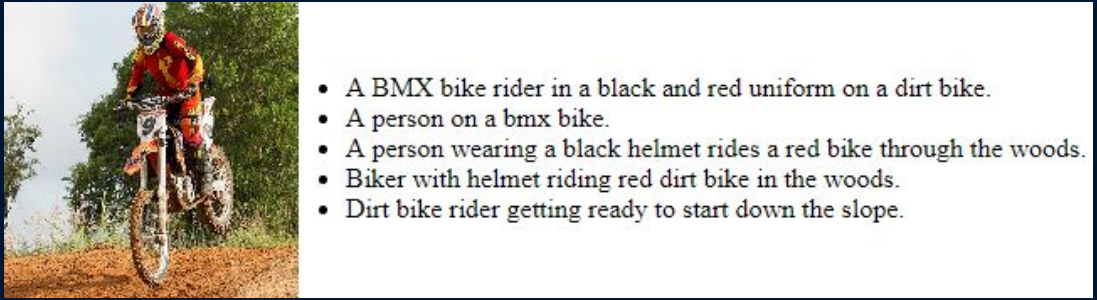
'a girl holding a horse lead behind a fire ',

'a man and girl and two horses are near a contained fire ',
'two people and two horses watching a fire '

DATASET DESCRIPTION

Flickr 8k dataset

It is a collection of 8,000 images with corresponding textual descriptions. The images cover a wide range of scenes, objects, and activities. Each image is accompanied by 5 different captions that describe the content of the image in different ways. This dataset is commonly used for training and evaluating image captioning models. It is a popular benchmark for testing the ability of these models to generate accurate and natural-sounding captions. The dataset is widely used in research. It provides a rich resource for developing and testing new approaches to image captioning.



- A BMX bike rider in a black and red uniform on a dirt bike.
- A person on a bmx bike.
- A person wearing a black helmet rides a red bike through the woods.
- Biker with helmet riding red dirt bike in the woods.
- Dirt bike rider getting ready to start down the slope.

RELATED WORKS

A lot of research has been done in this field.
Image captioning with attention: Incorporating an attention mechanism to improve the quality and relevance of generated captions.
Cross-modal retrieval: Retrieval of relevant captions or images using either modality as a query.
Multimodal fusion: Combining information from multiple modalities such as image, text, and audio to generate captions.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention Kelvin Xu et. al
Show, Attend and Tell is a neural image captioning model that uses an attention mechanism to generate captions for images. The model first encodes the input image into a set of feature vectors, then generates the caption by selectively attending to the relevant parts of the image during each step of the caption generation process.

METHODOLOGY

Preprocessing INPUT DATA

TEXT : We have applied a basic pre-processing techniques to clean the captions of the image such as removing non char words and punctuation.

IMAGE : To get the feature vectors of the images we have used different CNN pretrained models such as InceptionV3, VGG16, for different models.

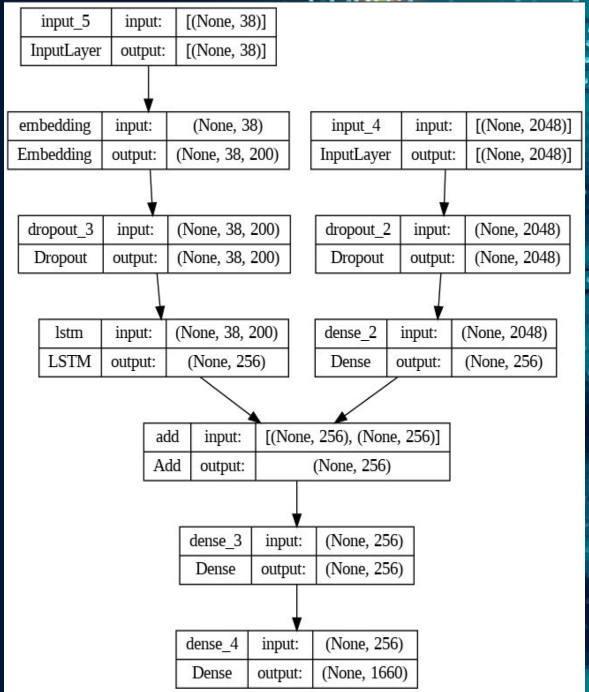
EMBEDDINGS: To get the embeddings of the captions we have used different embedding techniques such as Glove, tokenizers

METHODOLOGY

BaseLine
Our Base line model is a simple CNN + LSTM based model.

Where the CNN model is InceptionV3 model which takes images of size (229, 229) and generates their feature vector. Then this feature vector has been passed to the model, along with the text encoding to generate output sequence.

Text input encoding is passed through the LSTM then the both the modality inputs are being added and passed through a Dense layer followed by a softmax output layer on the vocabulary.



input_5 | input: | [(None, 38)]
InputLayer | output: | [(None, 38)]

embedding | input: | (None, 38)
Embedding | output: | (None, 38, 200)

input_4 | input: | [(None, 2048)]
InputLayer | output: | [(None, 2048)]

dropout_3 | input: | (None, 38, 200)
Dropout | output: | (None, 38, 200)

dropout_2 | input: | (None, 2048)
Dropout | output: | (None, 2048)

lstm | input: | (None, 38, 200)
LSTM | output: | (None, 256)

dense_2 | input: | (None, 2048)
Dense | output: | (None, 256)

add | input: | [(None, 256), (None, 256)]
Add | output: | (None, 256)

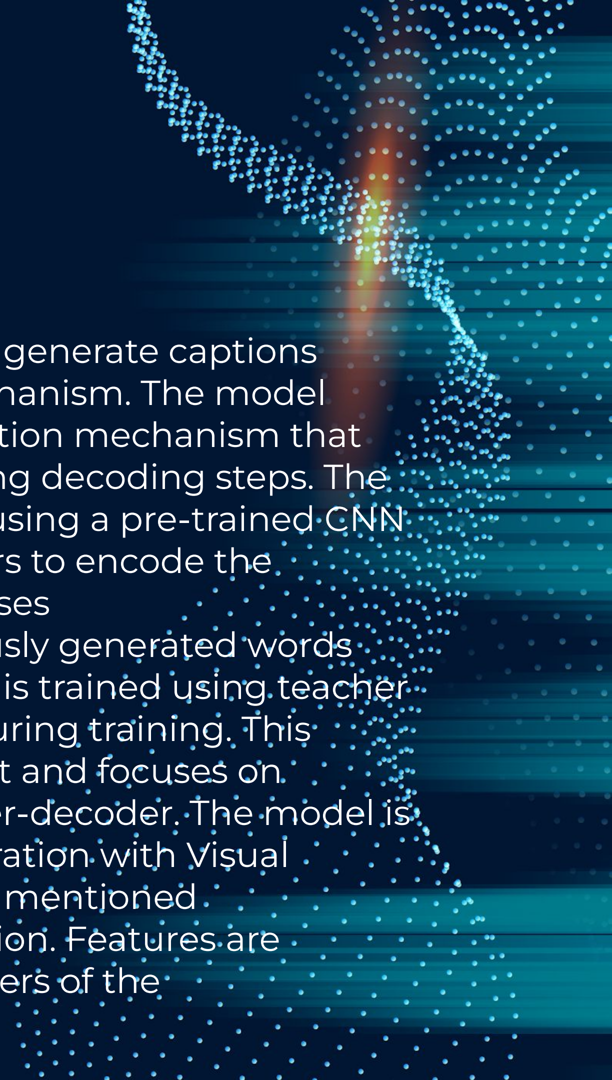
dense_3 | input: | (None, 256)
Dense | output: | (None, 256)

dense_4 | input: | (None, 256)
Dense | output: | (None, 1660)

METHODOLOGY

CNN + Visual Attention

"Image captioning with visual attention" demonstrates how to generate captions for images using a deep learning model with an attention mechanism. The model consists of an encoder-decoder architecture with a visual attention mechanism that allows the decoder to focus on different image regions in varying decoding steps. The model architecture in the tutorial first extracts image features using a pre-trained CNN (MobileNet), which is then passed through 2 self-attention layers to encode the features. Based on the Transformer architecture, the decoder uses cross-attention to attend to the image features and the previously generated words when generating the following word in the caption. The model is trained using teacher forcing, where the ground- truth captions are used as inputs during training. This model assumes that the pre-trained image encoder is sufficient and focuses on building the text decoder. This model uses a 2-layer Transformer-decoder. The model is inspired by ‘Show, Attend and Tell: Neural Image Caption Generation with Visual Attention’ but uses an updated 2-layer Transformer-decoder as mentioned above. The model architecture built works in the following fashion. Features are extracted from the image and passed to the cross-attention layers of the Transformer-decoder.



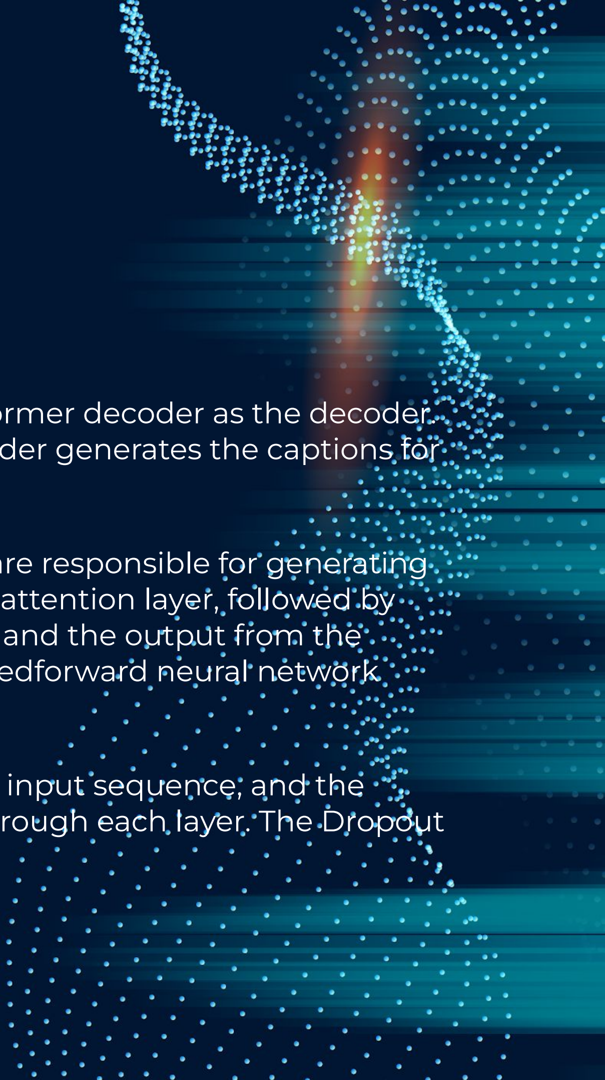
METHODOLOGY

CNN + Transformers (Multi - head attention)

The model uses an EfficientNetB0 CNN model as the encoder and a transformer decoder as the decoder. The encoder is used to extract features from the input image, and the decoder generates the captions for the input image.

The decoder consists of a stack of TransformerDecoderBlock layers, which are responsible for generating the output caption. Each block in the decoder consists of a multi-head self-attention layer, followed by another multi-head attention layer that takes the output from the encoder and the output from the previous attention layer as inputs. Finally, the output is passed through a feedforward neural network (FFN).

The PositionalEmbedding layer is used to add positional information to the input sequence, and the LayerNormalization layer is used to normalize the input before passing it through each layer. The Dropout layer is used for regularization to prevent overfitting.

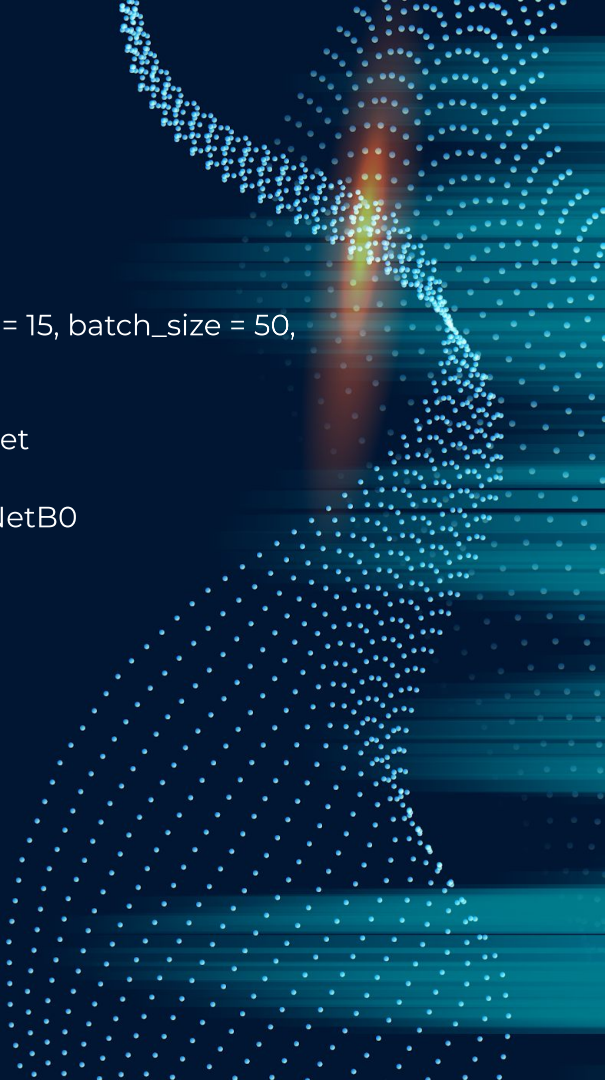


Hyperparameters

CNN + LSTM :- Glove embedding, cross_entropy, adam, softmax, relu, batch = 15, batch_size = 50, InceptionV3

CNN + VIsual Attn:- Masked loss, masked accuracy, adam, softmax, MobileNet

CNN + Multihead:- cross_entropy, early stoppings, adam, softmax, EfficientNetB0



EXPERIMENTS AND RESULTS

	BLEU - 1	BLEU - 2	ROUGE-L
BASELINE	0.55	0.36	0.53
VISUAL ATTENTION	0.54	0.33	0.56
MULTI HEAD	0.569	0.384	0.58

ANALYSIS

baseline1 uses a simple CNN (Inception-v3) and LSTM. Inception-v3 is a popular CNN architecture that is widely used for image classification. LSTM is a type of recurrent neural network that is commonly used for natural language processing tasks. The model uses Glove embeddings, which are pre-trained word embeddings.

Main Model uses a more advanced CNN architecture (EfficientNetB0) and a transformer decoder. EfficientNetB0 is a state-of-the-art CNN architecture that is designed to be computationally efficient while achieving high accuracy. The transformer decoder consists of a multi-head self-attention layer, which allows the model to focus on different parts of the input image and generate captions more effectively.

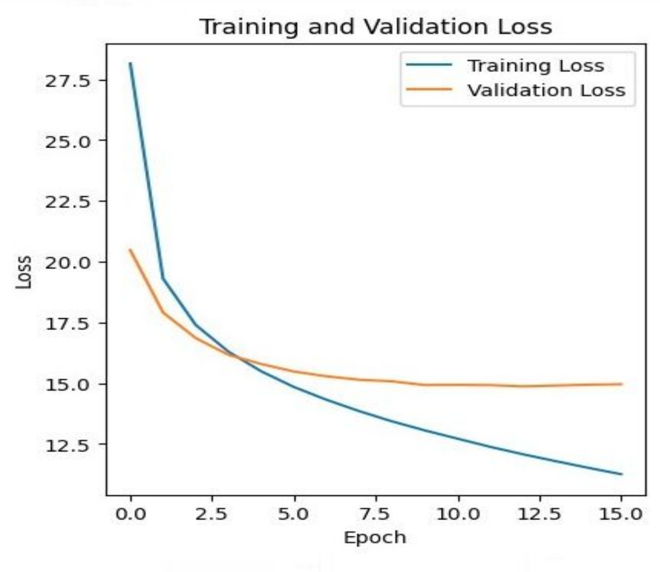
Baseline2 also uses a CNN architecture (MobileNet) and a transformer decoder, but it includes visual attention. Visual attention is a mechanism that allows the model to selectively focus on different parts of the input image when generating captions. This can help the model generate more accurate captions by focusing on the most important features in the image.

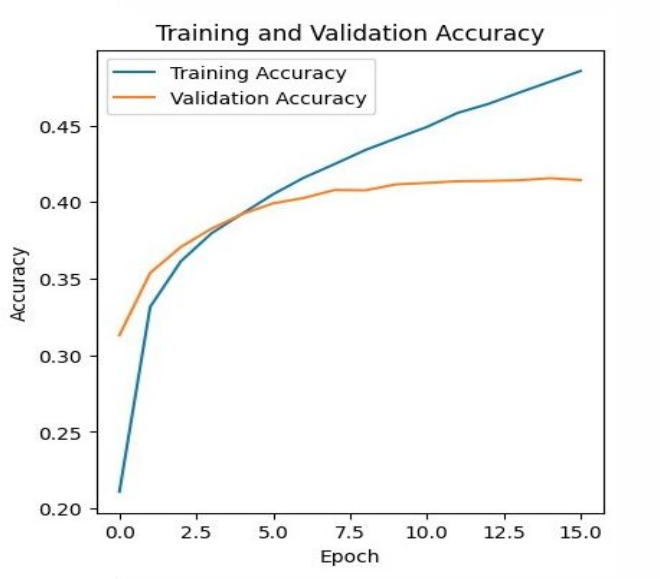
Overall, the main differences between these models are in the CNN architecture used and the type of decoder used for generating captions.

Amongst these, we got the best BLEU score on ‘Main model” due to its SOTA CNN+MultiHeadAttn approach.

LOSS CURVES

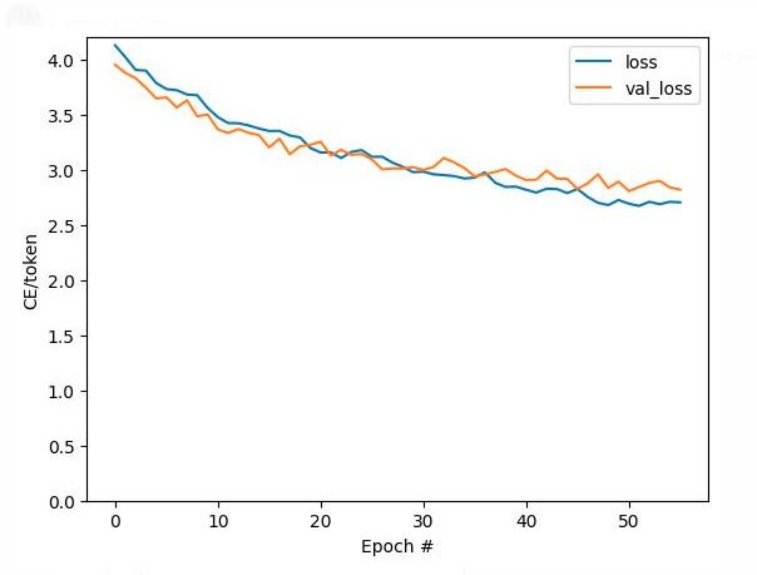
CNN + Multi-head Attn.

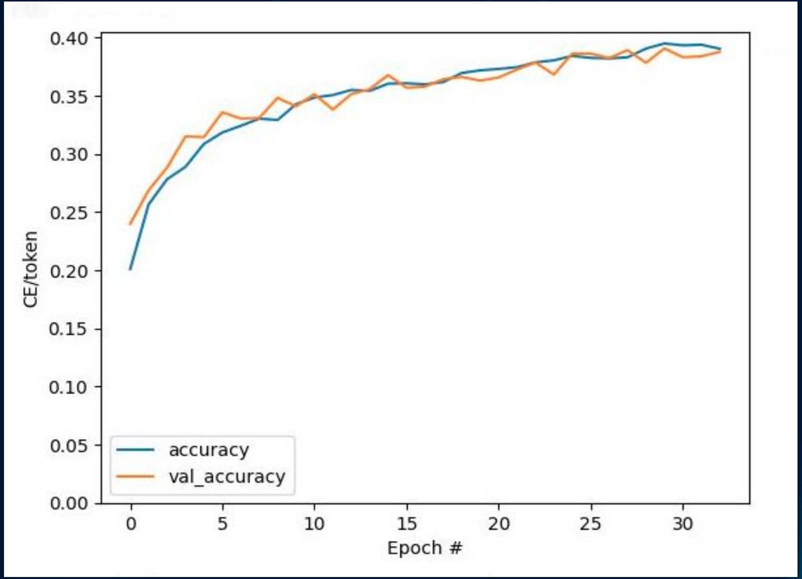




LOSS CURVES

CNN + Visual Attention





Aaditya Gupta

Contribution

CNN + Visual Attention

Harjeet Singh Yadav

contribution

CNN + LSTM Baseline

Harsh Vardhan Singh

contribution

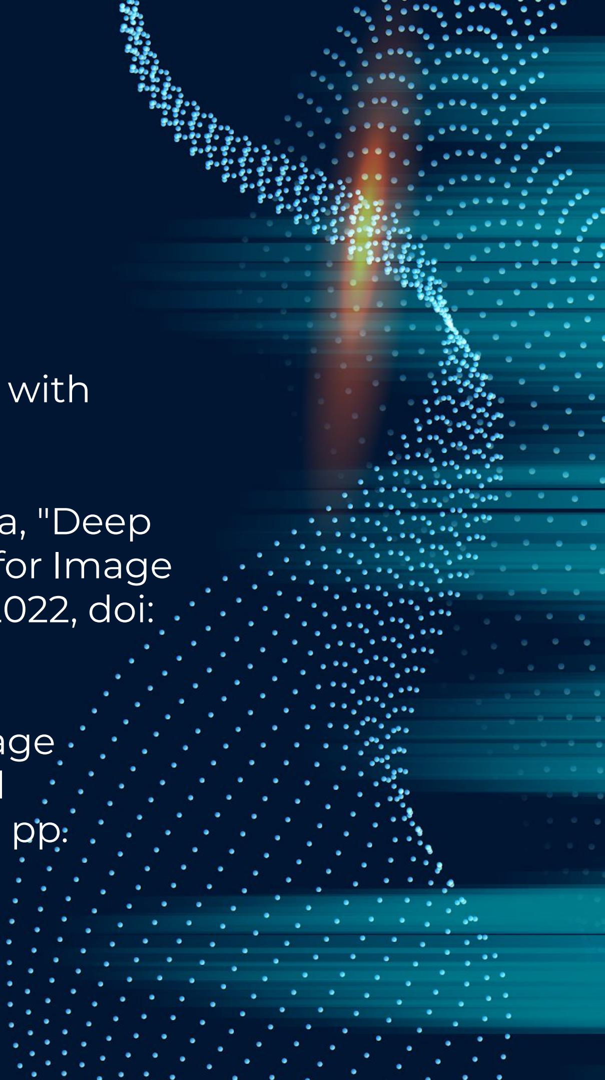
CNN + Transformer (Multi-head Attn)

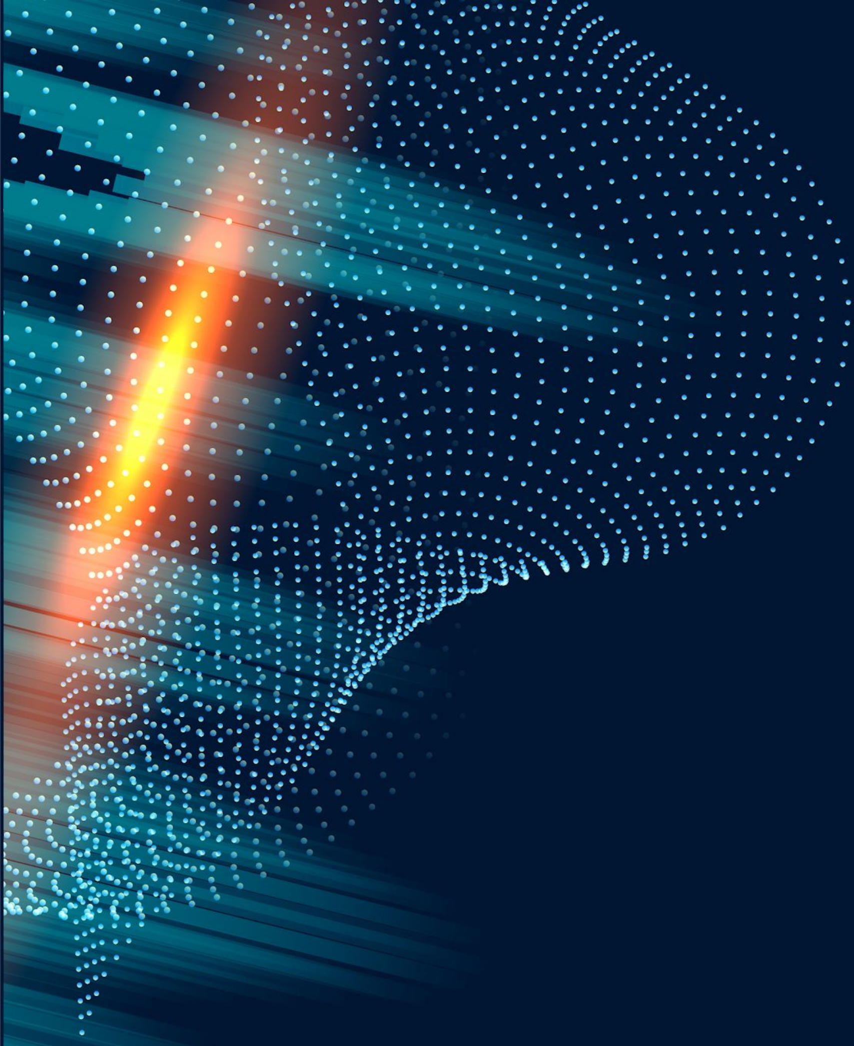
REFERENCES

[1] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, Kelvin Xu et. al

[2] R. Castro, I. Pineda, W. Lim and M. E. Morocho-Cayamcela, "Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks," in IEEE Access, vol. 10, pp. 33679-33694, 2022, doi: 10.1109/ACCESS.2022.3161428.

[3] Z. Zhang, Q. Wu, Y. Wang and F. Chen, "High-Quality Image Captioning With Fine-Grained and Semantic-Guided Visual Attention," in IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1681-1693, July 2019, doi: 10.1109/TMM.2018.2888822.





THANK YOU