The research thesis A comparative Analysis of two Image Captioning Novel Architectures
First of all , Image Captioning can be defined as transforming image into a text that describe it.

From a social perspective, the study aim to Enhance accessibility and Improve engagement for individuals with visual impairments. The picture example I provided explain how the captions can be heard by individuals with visual impairments.

- The study can also Improve image organization and search capabilities..for example , think of the apps used on the phone to organize our pictures. If you type in the search a word like food or vacation , the app will select all pictures related to that word. And that happens through image captioning and classification.

Scientifically speaking , I tried to briefly summarize the architectures used in Deep learning literature to address this image captioning task.

First , we have CNN + RNN (or lstm), I will explain the difference in next slide.

CNN is composed of encoder and decoder where the encoder extract image features and the decoder used for classification. For image captioning , we only use the decoder part by freezing layers of already trained network.

RNN or **recurrent** neural networks are networks that process data in an order manner. The gif picture I provided on the top is an example of translation where rnn process each word sequentially , and between each input we have a hidden state that preserve information that is used in the decoder part to produce outputs., however not all the preserved information is necessary.

So what is the difference between RNN and lstm? basically lstm is just an extended version of RNN that introduce gates such as input, forget, and output gates. These gates enabling LSTMs to selectively retain or discard information over time, which helps alleviate the vanishing gradient problem while its weights are being optimazed.. and capture long-range dependencies more effectively.
On the other hand, we have another architecture that is considered the state of art : cnn + transformer. The image on right top compare between the two architecture …as you can see , the both have cnn , and lstm process data sequentially to understand the syntax of a sentence while transformer process all inputs at once for each output it produce. This is good because it preserve global dependencies as well as semantic relationships through the use of attention mechanism that produce contextual embedding. I will explain that in the next slide.

traditional embeddings , which is numerical representation of a concept(like a word),…these embedding offer fixed representations regardless of context, which may like.. limit their ability to capture long dependencies …and meaningful semantic dimensions. …In contrast GPT/BERT embeddings provide contextual information by capturing word meanings based on context and.. capturing long-range dependencies between distant words. …Think of it as word being presented in several dimensions and the distance between related word or words with same/related meaning is close. For example , the word Netherlands , would be close to dutch , Amsterdam , while education and war may not necessary be presented close to each other….. This is because

they were trained on large-scale text datasets, that enable the embedding space to capture meaningful semantic dimensions.

So By combining CNN, Transformer (BERT or GPT), and LSTM components , we can address the limitations of existing methods , and …leverage the strengths of each…., simply that is what this research theory is based on .

So the Research Question is basically

- To what extent BERT and GPT enhance the contextual understanding and semantic coherence of image captions within a hybrid CNN-LSTM-Transformer architecture?

Sub-research questions:

- SQ1 What is the accuracy of the proposed hybrid architectures compared
- to baseline CNN-LSTM?
- SQ2 What are the differences in captioning accuracy and error type
- when using BERT versus GPT?
- SQ3 What captioning errors mistakes occur most frequently, and what
- the underlying reasons behind the recurring error patterns?

These questions will be address through the following research strategy