National College of Ireland

Advancements in Automated Image Captioning: A Comparative Study of Modern AI Models

MSc Research Project
Data Analytics

Shivakumar Patil
Student ID: 22144218

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shivakumar
Student ID:	22144218
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Arjun Chikkankod
Submission Due Date:	13/12/2023
Project Title:	Advancements in Automated Image Captioning: A Comparative Study of Modern AI Models
Word Count:	3800
Page Count:	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL Internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use another author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Shivakumar Patil
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on a computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Advancements in Automated Image Captioning: A Comparative Study of Modern AI Models

Shivakumar Patil
22144218

Abstract

The study presents a comprehensive study of full-sentence caption generation methods covering the overlap between visual content and natural language processing. Focused on Flickr dataset, study aims to explore recent approaches and compare 3 advanced methodologies including the combination of VGG-16 with LSTM, Vision Transformer (ViT) with GPT-2 and OpenAI's Contrastive Language–Image Pretraining (CLIP). Each approach is evaluated for its effectiveness in producing coherent and contextually relevant captions using BLEU-1 and BLEU-2 scores serving as the primary evaluation metrics and human evaluation. Additionally project briefly further studies potential NLP applications including trending generation, word based image search, translation and audio conversion. Eventually, this project aims to contribute this this latest evolving field of auto caption generation showcasing the capability and limitations of current approaches for future advancements in integrating visual and linguistic data processing and exploring potential use cases for these captions generated.

Keywords: Automated Image Captioning, VGG16 and LSTM, Vision Transformer (ViT) and GPT-2, CLIP (Contrastive Language–Image Pretraining), BLEU Scores, Natural Language Processing (NLP)

1 Introduction

1.1 Background

Image captioning has grown rapidly recently with the advance of deep learning and significant opportunities to improve image-capturing potential. It's one of the key areas in AI today bridging the gap between computer vision and natural language processing (NLP). Early approaches were primarily template-based and they have now become obsolete with the evolution of neural networks. Recent advances in Convolutional Neural Networks(CNN) have changed the field of image feature extraction rapidly and Recurrent Neural Networks(RNN) especially Long Term Memory (LSTM) networks have brought new opportunities in combining visual and generation of sequential data.

Example of caption generated using CLIP in this study and its potential use cases for visually impaired Figure 1

1.2 Research Question

Primary research question of this thesis is: " How can we leverage different deep learning models to generate multi-word captions and their effectiveness and its use cases?" The

Figure 1: Potential application: Generated caption translated and audio generated

thesis sets the below objectives to address the research question:

1. Evaluation: Study effectiveness of VGG16 with LSTM, ViT coupled with GPT-2 and CLIP models for full sentence/multi-word image captions generation.

2. Performance Comparison: Use metrics such as BLEU-1 and BLEU-2 scores to compare the performance of these 3 approaches along with human and human observation

3. Application Potential: Explore practical applications of these models in various domains such as translation and audio generation for the visually impaired, image trend analysis in the social media domain, sentiment analysis for the captions generated etc.

4. Limitations and Future Directions: Identify any limitations of current approaches of caption generated and recognize the future use case for the auto-generated captions

1.3 Document Structure

The thesis flow is as follows:

- Introduction: Sets the stage for the research background of image captioning, research question and objectives of this study
- Literature Review: Brief review of existing research in image captioning.
- Methodology: Details of methodologies used including VGG16 with LSTM, ViT with GPT-2 and CLIP for caption generation
- Experimental setup and data collection: Details of the Flickr dataset, image preprocessing and experimental setup

- Results and analysis: Evaluation of BLEU-1 and BLEU-2 scores including results of interpretation
- Conclusion and future work: Summarizes the study including its limitations & future work

2 Related Work

2.1 Overview:

Auto image captioning a combination of computer vision and natural language processing plays a key role in AI. Caption generation helps in a better understanding of visual content and has practical applications such as aiding visually impaired users, content indexing, and multimedia retrieval (Vinyals et al., 2015) [1].

While there has been significant work done recently in this field this literature review focuses on evaluating three prominent approaches:

1. **VGG16 and LSTM:**

   "Deep Visual-Semantic Alignments for Generating Image Descriptions" by Karpathy and Fei-Fei (2015) introduced combining CNN and RNN effectively to generate and retrieve descriptions of images (Karpathy and Fei-Fei, 2015) [2]

   VGG16 a deep CNN model was used to extract features from images (Simonyan and Zisserman, 2014) [6] along with LSTM for caption generation or sequential data.

2. **ViT and GPT-2:** Similarly combination of Vision Transformer (ViT) and GPT-2 was explored to how transformer architectures can be used in both visual perception and language modeling (Dosovitskiy et al., 2020; Radford et al., 2019) [3, 4].

3. **Pre-Trained Model-CLIP:**

CLIP model developed by OpenAI as outlined in Radford et al. (2021)[5] demonstrates remarkable zero-shot learning abilities which were developed by training on millions of diverse images paired with textual data. CLIP can be fine-tuned for generating captions, this ability was leveraged for multi-word captioning.

The review compares 3 caption generation methods. VGG16 and LSTM combination method show CNN and RNN synergy, ViT and GPT-2 show recent rise transformer architectures for multi-modal tasks and using pre-trained models like CLIP fine-tuned for caption generating. This study aims to evaluate their effectiveness.

2.2 VGG16 and LSTM

VGG16, a deep CNN model excels in extracting features from images (Simonyan & Zisserman, 2014) [6]. LSTM a type of recurrent neural network is well known for handling sequential data making it suitable for generating text (Hochreiter & Schmidhuber, 1997) [7].

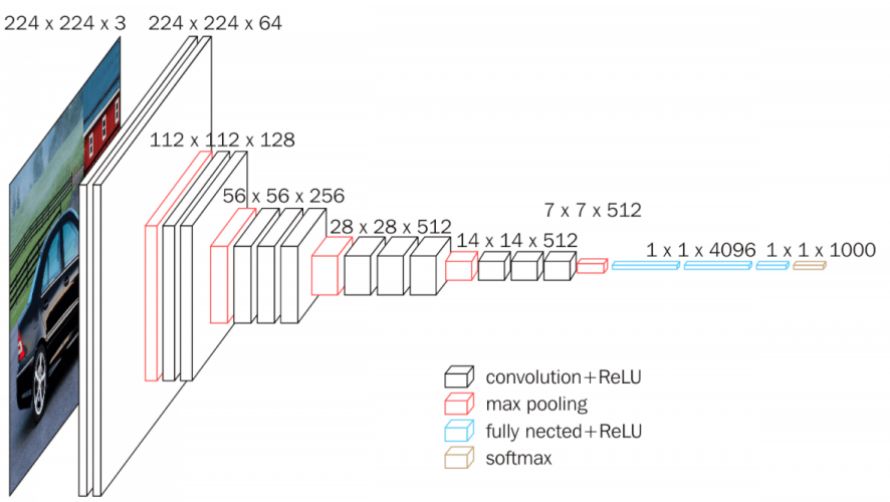

Figure 2: VGG16 [15]

2.2.1 Key Literature

1. Vinyals et al. (2015) [1]:

   - Paper Title: "Show and Tell: A Neural Image Caption Generator" (Vinyals et al., 2015) [1]
   - Contribution: The paper was key in the start of the neural network approach for image captioning, authors showed how CNN encodes an image and then use and then uses a recurrent neural network (RNN) to decode the CNN-generated vector into a descriptive sentence.
   - Impact: The model showed its ability to generate captions both accurately and linguistically coherently. A significant step forward in the field as it showed that a deep learning model could effectively bridge visual data with NLP.

2. Xu et al. (2015) [8]:

   - Paper Title: "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention"
   - Contribution: Building upon the CNN-LSTM framework, a visual attention mechanism was integrated into the image captioning model. This helped the model to focus of focus on different parts of the image, hence generating each word.
   - Impact: The attention concept was introduced in neural image captioning leading to a more focused and interpretable model. Since then they have become standard component in many image captioning models.

3. Lu et al. (2017) [12]:

   - Paper Title: "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning"

- Contribution: Proposed an adaptive attention model that used both visual information and previously generated words in generating captions. This gave an extra layer of flexibility and intelligence for cation generation

4. Rennie et al. (2017) [10]:

   - Paper Title: "Self-Critical Sequence Training for Image Captioning"
   - Contribution: This study introduced a new concept of 'self-critical sequence training'(SCST). SCST is a reinforcement learning approach where the model optimizes its performance based on the evaluation of the captions it generates against some predefined metrics this shifted the transition from traditional cross-entropy loss used, which helped.

2.3 Vision Transformer (ViT) and GPT-2

Combining Vision Transformer (ViT) and GPT-2 was a major recent development in image captioning. ViT proposed by Dosovitskiy et al. (2020), departs from traditional neural networks by using transformer architecture for processing image patches like words. GPT-2 by Radford et al. (2019) is a well-known text generation. Combining ViT and GPT-2 helped advance image captioning capabilities.

Key Literature

1. **Dosovitskiy et al. (2020): This paper set the** foundation for multimodal tasks like caption generation by showcasing ViT's ability in image classification.

2. **Radford et al. (2019):** GPT-2 paper showed its ability to understand and generate text, crucial for image captioning.

3. **Vasireddy et al. (2023)**[16] This study showed the integration of Vision Transformers (ViT) and GPT-2 for image captioning, which acted as the key inspiration for this project as one of the recent methodologies of image captioning.

ViT enhances visual understanding, while GPT-2 improves text generation, leading to more relevant and detailed captions. Vasireddy et al. (2023) [16] showed how this combination improved accuracy and context in caption generating however it requires high computational.

2.4 CLIP-Based Approach

OpenAI's Contrastive Language–Image Pretraining (CLIP) by Radford et al. [5] was the latest progress that revolutionized image captioning by training on diverse image-text pairs which made it better at generalization in its ability of captions.

Compared to VG16 - LSTM, ViT-GPT-2 has better advantages and a few limitations:
Advantages:

1. **Generalization and Contextual Relevance:** Since it is trained 400 million image-text pairs, it excels in diverse situations making it better than previous deep learning models that have a bias toward trained data sets and contextually relevant captions.

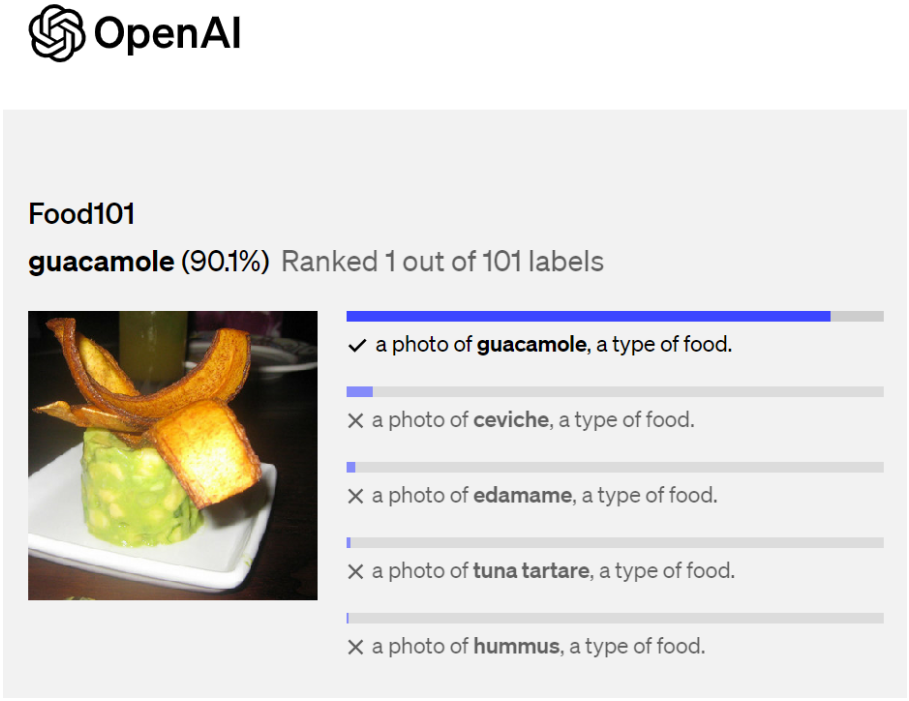

Figure 3: OpenAI's CLIP- Image Captions Demo [5]

2. **Zero-Shot Learning**: Zero-shot learning makes it different since it doesn't need to be trained for the task but might require fine-tuning. Like here, we give caption examples for fine-tuning the multi-word caption generation.

Limitations:

Precision: Its strength due to diverse training data also makes it less precise when the model is trained on a specific data set.

2.5 BLEU Scores

: BLEU (Bilingual Evaluation Understudy) scores have been standard for ML text-generated precision testing, Developed by Papineni et al. (2002) [14] BLEU scores compare the generated text vs reference text to compute a score.

3 Methodology

3.1 Data Collection and Preparation

- Dataset: Flickr8k and Flickr30K datasets used included 8000/30000 images with 5 diverse captions giving diverse captions data required for the training or fine-tuning the model.
- Image Prepossessing: Resizing (224x224 pixels) and normalization of images were done before feeding them into the training of (VGG16 - LSTM) and (ViT-GPT-2) models.
- Caption Processing: Captions were tokenized including padding or truncation to maintain a uniform length

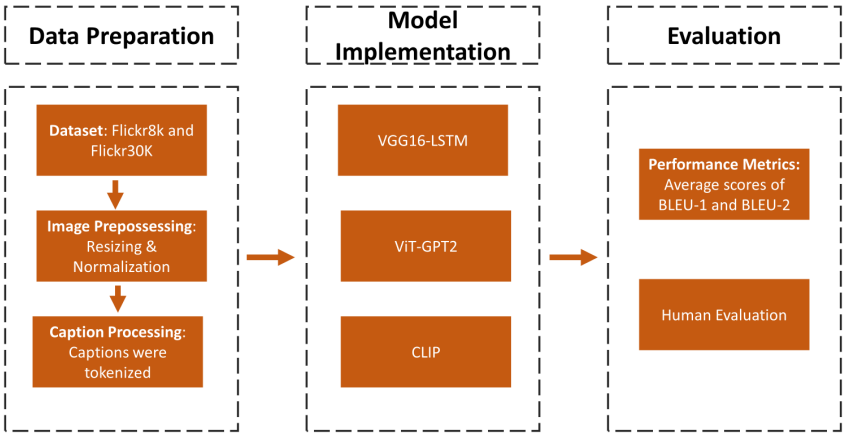

Figure 4: Methodology

3.2 Model Implementation and Setup

3.2.1 VGG16 and LSTM Setup:

- VGG16: Features were extracted from the penultimate layer VGG16, which will be used as the input for the LSTM.
- LSTM: LSTM network was appended that takes features from VGG16, these will be used to predict sequential data, in this case, sequence of words.

3.2.2 ViT and GPT-2 Setup:

- ViT: Vision Transformer for image analysis as used to extract features from the image
- GPT-2: GPT-2 was integrated into ViT for caption generation.

3.2.3 CLIP-Based Model Setup:

Fine-tuning of the CLIP model was done by feeding it the captions from the flikr8K dataset. These captions will be used as an example (Or fine-tuning is done using these caption examples) by the model built to generate captions.

3.3 Evaluation Methodology

- **Performance Metrics**: Average scores of BLEU-1 and BLEU-2 from the test/sample data were used to assess the performance and compare the models along with human evaluation.
- **Testing Procedure**: The flickr8k dataset was sampled for testing by either dividing as train and test data or if it was computationally heavy then test images were randomly sampled to 100-500 images and captions to calculate the average BLEU-1 and BLEU-2 score.

3.4 Training and Validation:

Training Process: ViT and GPT-2 and VGG16 and LSTM models were trained and optimum hyper-parameters were chosen. For the CLIP set-up, the model was fine-tuned using the captions from the Fliker dataset.

3.5 Result Interpretation:

Performance Comparison: By comparing the BLEU scores of each model setup, each model's accuracy was assessed in generating a caption.

Model Strengths and Weaknesses: Model strengths and limitations were assessed in situations where specific (Example: Medical diagnosis) vs general (Captions for visually impaired) image captions.

3.6 NLP Applications Overview

Several potential NLP applications were explored mostly by using the captions in the Flicker dataset itself to show possible use cases of the captions generated to set the ground for the future integration of model-generated captions once reliable options are generated.

3.6.1 Sentiment Analysis

(sentiment_analysis_on_the_captions_from_the_flickr8k_dataset.py): Sentiment analysis was done on the captions of images to analyse the distribution of sentiments of images (positive, neutral or negative).

3.6.2 Trending Topic Analysis (top_trending.py):

- **Objective**: Trending topics of the images were generated based on the captions
- **Procedure:** NLP techniques were used to extract key topics.
- **Application:** Can be used to analyze trending images uploaded in time period.

3.6.3 Searchable Database Creation (searchable_database.py):

- Objective: Searchable database of images based on their captions to extract images based on the word search
- Procedure: Word-based search to extract similar images based on matching with captions of the image in the database.
- Application: Efficient image search

3.6.4 Translation and Audio Conversion (clip_300_v3.py):

- Objective: Translate captions into different languages (Kannada) and audio.
- Procedure: Captions generated by the CLIP model were translated and audio conversion.

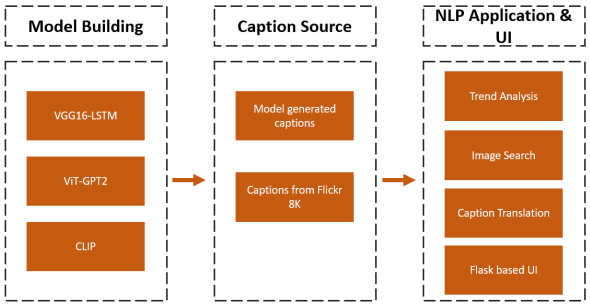

Figure 5: Project Design

- Application: To make users' image data more accessible to non-English speakers and the visually impaired.

4 Design Specification

4.1 Overview

Details of the architectures used are given for effective implementation:

4.2 1. VGG16 and LSTM Approach

- **Architecture**:
  - VGG16: A well-known CNN-based model for image recognition. It consists of 16 layers that have extracted image features.
  - LSTM (Long Short-Term Memory): An RNN was integrated to generate sequential data by taking VGG16 features.
- **Functionality**:
  - VGG16 processes images to extract feature vectors and these are used to input LSTM to generate captions.
  - LSTM takes both image feature vectors and also sequence of previously generated words.
- **Requirements**:
  - Pre-trained VGG16 model for feature extraction.
  - LSTM network
  - Training data (Images with captions): Flikr dataset

4.3 2. ViT and GPT-2 Approach

- **Architecture**:

  - Vision Transformer (ViT): A transformer-based model designed for Image analysis by converting images into patches and modelling them to find sequential relationships between those patches.
  - GPT-2: Transformer-based language model that is integrated with ViT to generate captions.

- **Functionality**:
  - ViT analyzes images and provides a contextual understanding of the visual content.
  - This understanding is then passed to GPT-2, which generates a corresponding caption based on the image context.

- **Requirements**:
  - Pre-trained ViT model for image analysis.
  - GPT-2 model for text generation.
  - Dataset with images

4.4 3. CLIP-Based Caption Generation

- **Architecture**: CLIP: Model trained on millions of internet image-text pairs
- **Functionality**: The model generates the most likely text description of the image
- **Requirements**:
  - Pre-trained CLIP model
  - Example descriptions/captions to fine-tune the CLIP mode for the caption description
  - Test dataset to check the accuracy

4.5 NLP Applications

4.5.1 Sentiment Analysis, Trending Topic Analysis, and Searchable Database Creation:

Integration of NLP techniques using captions generated by AI models or those sourced from the Flickr dataset for future integration when captioning accuracy becomes more reliable.

4.5.2 Translation and Audio Conversion:

In these direct integration is done with the captions generated and NLP applications were explored. Specifically leveraging captions generated by the CLIP model, these applications involve additional NLP processes for language translation and text-to-speech conversion.

Figure 6: Caption Translation and Audio in Kannada

5 Implementation

5.1 Overview

The project primarily focuses on 3 distinct latest AI-based approaches and their effectiveness in caption generation. Also, briefly touch upon potential NLP applications from the captions.

5.2 Model Implementations

1. **VGG16 and LSTM** integrated model was used for caption generation. The features extracted from the CNN-based VGG 16 model are fed into the LSTM network. **Output:** Multi-word captions for the image are generated

2. **ViT and GPT-2**: Captions were generated through Vision Transformer (ViT) for image processing which is then integrated with GPT-2. **Output**: Captions were closely aligned with the image content

3. **CLIP-Based Model:** Fine-tuned CLIP model was used for caption generation. **Output:** Diversly trained CLIP was able to generate accurate and context-aware captions

5.3 NLP Applications

- Conducted sentiment analysis, trending topic analysis and creation of a searchable image database using captions from the Flickr dataset but in future when models generate more reliable captions, these applications can be directly integrated with the model-generated captions.

- One such direct integration demo is shown here where CLIP-based model-generated captions are translated into another language (Kannada) and also audio conversion.

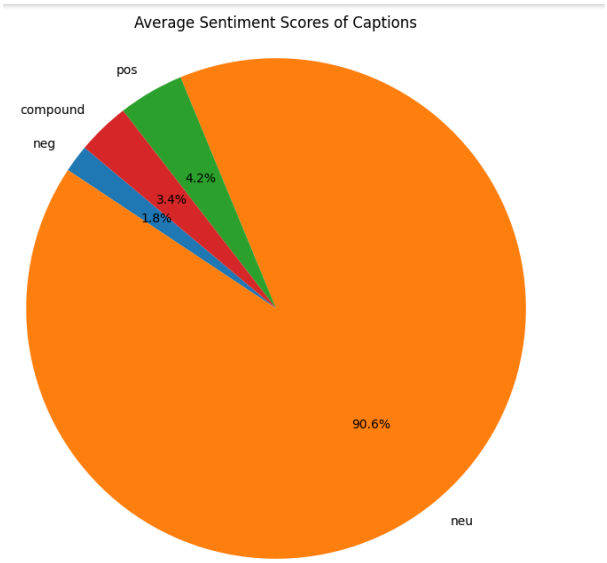

Figure 7: Image Sentiment Analysis

5.4 UI Extension :

Using Flask web framework, UI was created that allows users to upload images and captions are generated with CLIP as its backend.

5.5 Tools and Languages

- Python 3 is used throughout.
- Used libraries like TensorFlow, PyTorch and various NLP libraries

5.5.1 Outputs

- Three distinct caption generation models and their accuracy were studied and several NLP applications were explored.
- The output of this project includes generated captions, sentiment analysis results, identified trending topics, a searchable image database, translated and audio-converted captions and a UI interface.

6 Evaluation

6.1 VGG16-LSTM Approach

- Results: Average BLEU-1: 0.538675, BLEU-2: 0.289837.
- Analysis: The approach of using a combination of deep CNN and RNN, shows basic captioning proficiency. Moderate BLEU scores suggest limitations in capturing complex image details or generating, which could be due to LSTM's sequential nature, potentially limiting creativity in language generation.

Double-click (or enter) to edit

```
[ ] generate_caption("1313987366.jpg")
```

```
---------------------Actual---------------------
startseq woman in dress is holding camera next to another woman who is laughing endseq
startseq two young woman are standing in living room laughing and smiling endseq
startseq these two women are having fun while taking pictures endseq
startseq two women laughing at cat walking on the furniture endseq
startseq two young women clap their hands in laughter endseq
--------------------Predicted--------------------
startseq woman in black dress is dancing with her friend endseq
```

Figure 8: VGG16 LSTM Model Captions

- Implications: The model may need further work to capture complex scenarios/ diverse linguistic expressions.

6.2 ViT-GPT2 Approach

- **Results:** Average BLEU-1: 0.516098, BLEU-2: 0.314810
- **Analysis**: This transformer-based image model improved the ability to capture complex sentence structures as indicated by the BLEU-2 score. However, performance was comparable to VGG16-LSTM.
- **Implications:** While further fine-tuning is required but has the potential to generate more linguistically rich captions.

6.3 CLIP Approach

- **Results:** Average BLEU-1 score of 0.587136 and BLEU-2 score of 0.399473 for the CLIP approach.
- Analysis: CLIP gives the best score compared to other approaches, showcasing the better ability to generate quality captions, indicating CLIP's ability.

```
['A black and white dog is running through a grassy field..',
 'A black and white dog is running through a grassy area..',
 'A black and white dog is running through a field... and']
```

Figure 9: ViT-GPT2

```
Caption: A little girl in a pink dress going into a wooden cabin .
```

Figure 10: CLIP Model Caption

6.4 Discussion

6.4.1 Analysis of Findings:

This study showed varying levels of effectiveness among 3 approaches studied, with the CLIP model outperforming other approaches based on the BLEU score.

6.4.2 Critique of Experiments:

BLEU scores alone may not be sufficient for measuring the model's accuracy, better metrics may be required and VGG16-LSTM failed in generating diverse linguistic expressions.

6.4.3 Suggested Improvements:

Additional evaluation metrics and integrating attention mechanisms in VGG16-LSTM might enhance its ability.

6.4.4 Contextualization with Literature:

Findings align with the existing literature, with the latest approach, CLIP was better at getting captions for general captioning applications.

7 Conclusion and Future Work

Project provided key insights into auto image captioning space by showcasing the effectiveness of 3 latest AI based approaches: VGG16 ¿¿ LSTM, ViT ¿¿ GPT-2 and CLIP. Results of the study align with the existing literature review done in the beginning showcasing OpneAI's CLIP model's ability to generate accurate and contextually relevant captions for images in general cases.

Study is limited to the reliance on BLEU scores for model performance measurement and hence future work could explore more holistic metrics. And integrate the NLP application directly from the model generated captions.

Future Work:

- **Advanced Evaluation Metrics:** Use more comprehensive evaluation metrics to assess caption quality better including METEOR, CIDEr and ROUGE.
- **Commercial Applications:** Study for more potential practical applications of automated image captioning and possible business use cases for possible revenue generation.
- **Real-Time Captioning**

8 Acknowledgement

I would like to thank professor Arjun Chikkankod for supporting me and guiding me throughout this process. His expertise and support were key to completing this research project.

9 References

1. Vinyals, O., Toshev, A., Bengio, S., and Erhan, D., 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

2. Karpathy, A. and Fei-Fei, L., 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... and Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I., 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*.

5. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... and Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

6. Simonyan, K. and Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

7. Hochreiter, S. and Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation*.

8. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y., 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning* (pp. 2048-2057).

9. Lu, J., Xiong, C., Parikh, D., and Socher, R., 2017. 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* .

10. Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V., 2017. Self-Critical Sequence Training for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* .

11. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I., 2020. Generative Pretraining from Pixels. In *International Conference on Machine Learning*.

12. Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). 'Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

13. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). 'Self-Critical Sequence Training for Image Captioning'.

14. Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). 'BLEU: a Method for Automatic Evaluation of Machine Translation'.

15. Kamilaris, A. and Prenafeta-Boldú, F.X. (2018). A review of the use of convolutional neural networks in agriculture. The Journal of Agricultural Science.

16. Indrani Vasireddy, G.Hima Bindu and Ratnamala. B (2023). Transformative Fusion: Vision Transformers and GPT-2 Unleashing New Frontiers in Image Captioning within Image Processing. International journal of innovative research in engineering and management.