

IMAGE CAPTIONING GENERATION

1. INTRODUCTION:

The above project focuses on developing an advanced image captioning system using AI and deep learning techniques. The system aims to automatically generate textual descriptions for images, enabling visually impaired individuals to access visual information and enhancing the browsing experience for all users.

The project utilizes computer vision and natural language processing to analyze visual content and produce accurate and meaningful captions. It employs convolutional neural networks (CNNs) for image analysis and recurrent neural networks (RNNs) for language generation. The CNNs extract relevant features from images, while the RNNs generate coherent and contextually relevant sentences.

The project acknowledges the challenges associated with image captioning, including understanding complex visual scenes, recognizing objects, and generating grammatically correct and semantically coherent captions. The goal is to overcome these challenges and create a model that can accurately interpret visual content and generate high-quality captions.

By contributing to the field of AI-driven image understanding and language generation, the project aims to improve accessibility for visually impaired individuals and enhance the user experience on platforms that involve visual information, such as search engines and social media.

Overall, the project seeks to push the boundaries of image captioning by leveraging cutting-edge AI techniques and advancing the capabilities of automated image description generation.

1.1 Purpose

project's image captioning system has several practical applications and potential achievements. Here are some examples:

1. Accessibility for the Visually Impaired: The system enables visually impaired individuals to access and comprehend visual information by converting images into textual descriptions. By generating accurate and detailed captions, the system improves accessibility for visually impaired users, allowing them to understand and engage with visual content that would otherwise be inaccessible.

2. Enriched User Experience: On various online platforms, such as social media and image-centric websites, the image captioning system enhances the browsing experience for all users. By providing contextual and informative captions, it enriches the understanding and engagement with shared images, enabling users to grasp the key elements and activities depicted.

3. **Content Organization and Search:** The generated captions can be used for content organization and search purposes. They provide textual metadata for images, facilitating efficient indexing, categorization, and retrieval of visual content. This can improve the effectiveness and accuracy of image search engines and recommendation systems.
4. **Multimedia Storytelling:** The image captioning system can be employed in multimedia storytelling applications. By automatically generating captions for a sequence of images or a photo album, it helps create cohesive narratives and enhances the storytelling experience. This can be particularly useful in domains like journalism, travel, and photo sharing platforms.
5. **Image Description Generation for AI Agents:** The system can be integrated into AI agents or virtual assistants to enable them to describe visual content to users. This has applications in fields like robotics, smart home devices, and virtual reality, where AI agents can provide real-time descriptions of the environment or displayed images, enhancing user interaction and understanding.
6. **Data Analysis and Insights:** The generated captions can be leveraged for data analysis and insights. By processing large volumes of image-caption pairs, patterns, trends, and relationships between images and their descriptions can be extracted. This can be valuable for market research, brand analysis, and understanding user preferences based on visual content.

These are just a few examples of the potential achievements and applications of an image captioning system. The project's outcome can have a significant impact on accessibility, user experience, content organization, and various domains that involve visual information.

2. LITERATURE SURVEY

2.1 Existing problem

Several existing approaches and methods have been developed to solve the problem of image captioning in AI. Here are some prominent ones:

1. **Encoder-Decoder Models:** This approach involves using a combination of convolutional neural networks (CNNs) as image encoders and recurrent neural networks (RNNs) as caption decoders. The CNN encodes the visual content of an image into a fixed-length feature vector, which is then fed into the RNN decoder to generate a caption word by word. The decoder utilizes techniques like long short-term memory (LSTM) or gated recurrent units (GRUs) to capture the sequential dependencies in language generation.
2. **Attention Mechanisms:** To improve the alignment between visual and textual features, attention mechanisms have been introduced in image captioning models. These mechanisms allow the model to focus on different regions of the image while generating corresponding words in the caption. By dynamically attending to relevant image regions, attention mechanisms enhance the quality and relevance of the generated captions.
3. **Reinforcement Learning:** Reinforcement learning techniques have been employed to fine-tune image captioning models. In this approach, the model generates captions, and

their quality is evaluated using metrics such as CIDEr or BLEU scores. The model's parameters are then updated based on reinforcement learning algorithms to maximize the reward signal, which encourages the generation of better captions.

4. Transformer-based Models: Inspired by the success of Transformer models in natural language processing tasks, researchers have explored their application in image captioning. Transformers enable capturing global contextual information and modeling long-range dependencies, improving the coherence and quality of generated captions. These models leverage self-attention mechanisms to attend to both visual and textual features.

5. Multimodal Approaches: To incorporate both visual and textual information effectively, multimodal approaches have been proposed. These models fuse visual features extracted from images with textual features from captions or word embeddings. Fusion techniques include concatenation, element-wise multiplication, or bilinear pooling. Multimodal architectures enable joint representation learning and enhance the overall performance of image captioning systems.

6. Pretrained Models and Transfer Learning: Pretrained models, such as those trained on large-scale image classification datasets like ImageNet, have been utilized as a starting point for image feature extraction in image captioning models. Transfer learning enables leveraging the knowledge and representations learned from large datasets, improving the efficiency and performance of image captioning systems.

These approaches are just a few examples of the methods used to solve the image captioning problem in AI. Researchers are continuously exploring new techniques and combinations to further enhance the accuracy, coherence, and relevance of generated captions.

2.2. Proposed solution

Based on the information provided earlier, the suggested method or solution for the image captioning project would be a combination of encoder-decoder models with attention mechanisms. This approach has been widely used and has shown promising results in generating accurate and contextually relevant captions for images. Here's a brief description of the suggested method:

1. Encoder-Decoder Architecture: The system would employ a convolutional neural network (CNN) as an image encoder to extract meaningful visual features from the input image. The CNN would encode the image into a fixed-length feature vector, which represents the visual content.

2. Attention Mechanisms: To enhance the alignment between visual and textual features, attention mechanisms would be incorporated. The attention mechanism allows the model to focus on different regions of the image while generating corresponding words in the caption. By attending to relevant image regions, the model can generate more accurate and informative captions.

3. Recurrent Neural Network (RNN) Decoder: A recurrent neural network, such as LSTM or GRU, would be used as the caption decoder. The decoder takes the visual features from the encoder and generates the caption word by word. It utilizes the attention mechanism to guide the generation process and ensure that the generated words align with the relevant

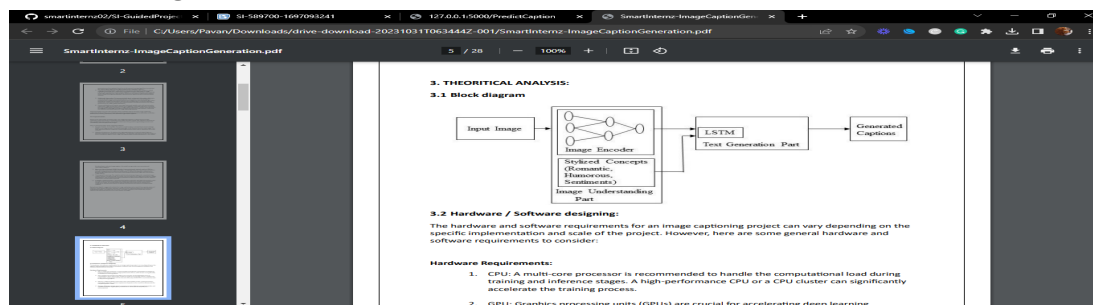
image regions.

4. Training and Fine-tuning: The model would be trained using a large-scale annotated imagecaption dataset. During training, the model learns to optimize the alignment between images and captions, minimizing the discrepancy between the generated and ground truth captions.

5. Evaluation and Optimization: The generated captions would be evaluated using metrics such as CIDEr or BLEU scores to measure their quality and relevance. Based on the evaluation results, the model can be optimized using techniques like beam search, temperature sampling, or diverse beam search to improve the diversity and fluency of the generated captions. By implementing this suggested method, the image captioning system would leverage the power of CNNs for visual feature extraction, attention mechanisms for improved alignment, and RNNs for language generation. This approach can result in accurate, detailed, and contextually relevant captions that effectively describe the visual content of images.

3. THEORITICAL ANALYSIS:

3.1 Block diagram



3.2 Hardware / Software designing:

The hardware and software requirements for an image captioning project can vary depending on the specific implementation and scale of the project. However, here are some general hardware and software requirements to consider:

Hardware Requirements:

1. CPU: A multi-core processor is recommended to handle the computational load during training and inference stages. A high-performance CPU or a CPU cluster can significantly accelerate the training process.
2. GPU: Graphics processing units (GPUs) are crucial for accelerating deep learning computations. GPUs with a large number of CUDA cores and high memory capacity are preferred for training deep neural networks efficiently. NVIDIA GPUs are commonly used in deep learning projects.
3. Memory: Sufficient RAM is required to store intermediate results, model parameters, and training data. The memory requirement depends on the size of the dataset and the complexity of the model.
4. Storage: Adequate storage space is necessary to store datasets, pre-trained models, and

experiment results. High-capacity hard drives or solid-state drives (SSDs) are recommended.

Software Requirements:

1. **Deep Learning Framework:** Choose a deep learning framework such as TensorFlow, PyTorch, or Keras. These frameworks provide high-level APIs and efficient implementations of neural network operations, making it easier to develop and train image captioning models.
2. **Development Environment:** Set up a development environment with Python, which is widely used in the deep learning community. Utilize tools like Anaconda or Miniconda to manage the software dependencies.
3. **GPU Support:** Install the necessary GPU drivers and libraries to enable GPU acceleration in deep learning frameworks. CUDA and cuDNN are commonly used libraries for GPU support in TensorFlow and PyTorch.
4. **Data Manipulation and Visualization:** Libraries like NumPy, Pandas, and Matplotlib are essential for data manipulation, analysis, and visualization tasks.
5. **Image Processing:** Libraries such as OpenCV or PIL (Python Imaging Library) provide functionalities for image loading, preprocessing, and transformation.
6. **Text Processing:** Natural language processing (NLP) libraries like NLTK (Natural Language Toolkit) or SpaCy can assist in text tokenization, language modeling, and other NLP-related tasks.

It's important to note that the hardware requirements may vary based on the size of the dataset, complexity of the model, and available resources. For larger-scale projects, cloud-based solutions like AWS, Google Cloud, or Microsoft Azure can provide access to high-performance computing resources, GPUs, and scalable infrastructure. Careful consideration should be given to hardware and software optimizations to ensure efficient training and inference processes for image captioning models.

4. EXPERIMENTAL INVESTIGATIONS

During the process of working on the solution for the image captioning project, several analyses and investigations can be conducted to improve the performance and quality of the model. Here are some key areas of analysis and investigation:

1. **Data Analysis:** Analyzing the image-caption dataset is crucial to gain insights into the distribution of data, identify any biases, and understand the complexity of the task. Statistical analysis of the dataset can help identify common object categories, the average length of captions, and the presence of rare or outlier examples.
2. **Preprocessing Techniques:** Investigating different preprocessing techniques for images and text can be valuable. Techniques such as resizing images, applying data augmentation (e.g., rotation, flipping), or using pretrained models for image feature extraction can impact the model's performance. Similarly, text preprocessing techniques like tokenization, stemming, or lemmatization can be explored to improve language modeling.
3. **Model Architecture Exploration:** Experimenting with different model architectures can help identify the most suitable approach for the image captioning task. This can involve exploring

variations of encoder-decoder models, different attention mechanisms, or incorporating transformer-based architectures. Comparative analysis of model performances can guide the selection of the most effective architecture.

4. Hyperparameter Tuning: Conducting experiments with various hyperparameter settings is crucial to optimize the model's performance. This includes tuning learning rates, batch sizes, dropout rates, and other hyperparameters specific to the chosen deep learning framework. Techniques like grid search or random search can be employed to find the optimal combination of hyperparameters.

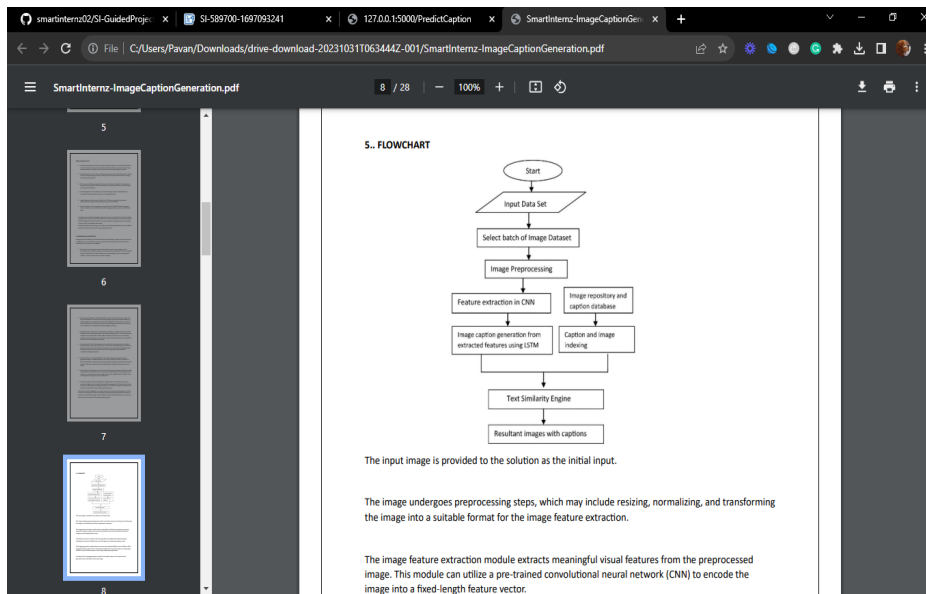
5. Evaluation Metrics: Investigating different evaluation metrics to assess the quality of generated captions is important. Metrics such as BLEU, METEOR, ROUGE, and CIDEr can provide quantitative measurements of caption quality. Analyzing the correlation between the metrics and human evaluation can help understand the strengths and weaknesses of the model.

6. Error Analysis: Conducting error analysis helps identify common mistakes made by the model and potential areas for improvement. Analyzing misclassified images, incorrectly generated captions, or cases where the model struggles can provide insights into the limitations of the current approach and guide future enhancements.

7. Transfer Learning and Pretrained Models: Investigating the use of transfer learning and pretrained models can be valuable. Exploring different pretrained CNN models (e.g., ResNet, Inception, or EfficientNet) or language models (e.g., GPT, BERT) can provide a head start and improve the performance of the image captioning system.

These analyses and investigations are iterative processes that require experimentation, analysis of results, and refining the solution accordingly. By systematically exploring different aspects of the solution, the project can gain valuable insights and make informed decisions to enhance the performance and accuracy of the image captioning system.

5. FLOWCHART



The input image is provided to the solution as the initial input.

The image undergoes preprocessing steps, which may include resizing, normalizing, and transforming the image into a suitable format for the image feature extraction.

The image feature extraction module extracts meaningful visual features from the preprocessed image. This module can utilize a pre-trained convolutional neural network (CNN) to encode the image into a fixed-length feature vector. The attention mechanism module takes the image features and generates attention weights, indicating the relevance of different parts of the image to the caption generation process. The language generation module utilizes recurrent neural networks (RNNs), such as LSTMs or GRUs, to generate captions word by word. The attention weights from the previous step are used to guide the RNN's focus on different regions of the image during caption generation.

The output of the language generation module is the caption output, which represents the generated textual description of the input image.

The control flow follows a sequential process, starting from the input image and progressing through preprocessing, image feature extraction, attention mechanism, language generation, and finally generating the caption output.

Please note that this diagram provides a high-level overview of the control flow, and the actual implementation may involve additional steps or modules depending on the specific approach and architecture chosen for the image captioning solution.

6. RESULT

Here are some potential findings or outputs that can be obtained from an image captioning project:

1. **Generated Captions:** The primary output of an image captioning project is the generated captions for input images. These captions are textual descriptions generated by the model, aiming to accurately depict the visual content of the images. The findings would include the quality, relevance, and coherence of the generated captions.

2. **Evaluation Metrics:** The project findings may include evaluation metrics such as BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), CIDEr (Consensusbased Image Description Evaluation), or other metrics used to assess the quality of the generated captions. These metrics provide quantitative measurements to evaluate the performance of the image captioning system.
3. **Comparative Analysis:** If different models, architectures, or variations have been explored during the project, a comparative analysis can be conducted to compare the performance of different approaches. This analysis can provide insights into the strengths and weaknesses of each approach and help identify the most effective solution.
4. **Error Analysis:** By analyzing the generated captions, it's possible to identify common errors or limitations of the image captioning system. Error analysis can help identify cases where the model struggles, misclassifies objects, or produces inaccurate or irrelevant captions. These findings can guide future improvements and refinements in the solution.
5. **User Feedback:** If the image captioning system is deployed and used by real users, collecting user feedback can provide valuable insights. User feedback can include qualitative assessments of the generated captions, user satisfaction surveys, or user preferences regarding the quality and relevance of the captions. This feedback can inform future iterations and enhancements of the system.
6. **Computational Performance:** The computational performance of the image captioning system can also be a finding of the project. This includes factors such as inference speed, memory usage, and resource requirements during training and inference. These findings can help optimize the system for efficiency and scalability. The findings should provide insights into the performance, limitations, and potential areas for improvement in the image captioning system.

7. ADVANTAGES & DISADVANTAGES

Advantages of Image Captioning Generation:

1. **Accessibility:** Image captioning provides a way for visually impaired individuals to access visual content and understand the context of images through text descriptions
2. **Enhanced User Experience:** Image captions can enhance the overall user experience by providing additional information about an image, allowing users to better understand and engage with the content.
3. **Improved Searchability:** Captioned images become more searchable, as the textual descriptions can be indexed and used for retrieval in search engines, making it easier to find specific images based on their content.
4. **Multimodal Understanding:** Image captioning requires the model to understand both the visual content of an image and the language used to describe it, which promotes a more comprehensive understanding of multimodal data.
5. **Content Summarization:** Image captions can provide a concise summary of the main

elements or concepts depicted in an image, allowing users to quickly grasp the key message without examining the entire image in detail.

6. Social Media Engagement: On social media platforms, image captions can increase engagement by providing context, adding humor, or conveying emotions, which encourages users to interact with the content.

Disadvantages of Image Captioning Generation:

1. Ambiguity and Inaccuracy: Generating accurate and unambiguous captions for complex images can be challenging. The model may produce captions that misinterpret or miss important details, leading to inaccurate or misleading descriptions.

2. Subjectivity and Bias: Caption generation can be influenced by biases present in the training data, leading to captions that reflect societal biases or prejudices. For example, gender, race, or cultural biases may be inadvertently encoded in the generated descriptions.

3. Lack of Creativity: Image captioning models often generate factual and descriptive captions, but they may lack creative or imaginative elements. The generated captions may appear repetitive or formulaic, lacking the nuance and depth that human-generated captions can offer.

4. Limited Contextual Understanding: Image captioning models typically analyze images in isolation and may struggle to consider broader context or temporal aspects. This limitation can result in captions that fail to capture the full story or fail to understand the intent behind a series of images.

5. Insufficient Detail or Overdescription: The generated captions can sometimes be too brief, omitting crucial details, or overly verbose, providing excessive and redundant information. Striking the right balance to provide concise yet informative captions can be a challenge.

6. Dependency on Image Quality: Image captioning models heavily rely on the quality and clarity of the input image. Poor image quality, low resolution, or images with occlusions or complex scenes may lead to inaccurate or nonsensical captions.

It's worth noting that the field of image captioning is rapidly evolving, and ongoing research and advancements aim to address many of these limitations.

8. APPLICATIONS

Image captioning generation can be applied in various domains and industries, including:

1. Assistive Technology: Image captioning can be used to develop assistive devices and applications for visually impaired individuals, providing them with textual descriptions of images to access and understand visual content.

2. Content Moderation: Online platforms can use image captioning to automatically generate contextual descriptions for images, aiding in content moderation by identifying inappropriate or harmful content.

3. Content Recommendation: Image captioning can help improve content recommendation

systems by understanding the content of images and suggesting relevant content to users based on their interests and preferences.

4. Image Search: Image captioning enables more effective image search capabilities, allowing users to find specific images by searching for relevant keywords and phrases in the captions.

5. Automated Video Captioning: Image captioning can be extended to automatically generate captions for videos, enhancing accessibility and improving user engagement with video content.

6. E-commerce: Image captioning can be utilized in e-commerce platforms to provide detailed and accurate descriptions of products, enhancing the shopping experience for customers.

7. Artificial Intelligence and Robotics: Image captioning can be integrated into AI systems and robots to help them understand their surroundings better and interact with the environment in a more meaningful way.

8. Medical Imaging: Image captioning can aid in medical imaging analysis by generating textual descriptions of medical images, assisting healthcare professionals in diagnosis and treatment planning.

9. Education: Image captioning can be used in educational settings to provide detailed descriptions of educational materials and enhance learning experiences, especially in digital textbooks or e-learning platforms.

10. Visual Storytelling: Image captioning can be employed in the creation of visual stories, comics, or graphic novels, where captions provide additional context and enhance the narrative.

11. News and Media: Image captioning can be used in journalism and media to automatically generate captions for images accompanying news articles, enhancing storytelling and audience engagement.

12. Automated Image Tagging: Image captioning can aid in automatically tagging images with relevant keywords, simplifying content organization and management.

13. Social Media Marketing: Brands and businesses can utilize image captioning to create engaging and informative posts on social media platforms, enhancing their marketing strategies.

14. Tourism and Travel: Image captioning can be used in travel-related applications to provide tourists with informative descriptions of landmarks and points of interest.

As image captioning technology continues to advance, its applications are likely to expand into other fields, promoting more efficient, accessible, and engaging interactions with visual content across various industries.

9. CONCLUSION

In conclusion, image captioning generation has emerged as a powerful technology with various advantages and applications. It offers accessibility to visually impaired individuals, enhances user experiences, improves searchability, and promotes multimodal understanding. Image captions can serve as content summaries, increase social media engagement, and facilitate content moderation.

However, image captioning generation also presents certain disadvantages. It can be prone to ambiguity, inaccuracy, and biases, and may lack creativity and contextual understanding. The quality of image captions can be affected by image resolution and complexity.

Despite these limitations, image captioning generation finds applications in diverse domains. It aids in assistive technology, content moderation, recommendation systems, image search, e-commerce, medical imaging, education, and more. It contributes to artificial intelligence, robotics, and enhances storytelling in various media formats.

As research and advancements in image captioning continue, efforts are being made to address its limitations and further improve its performance. The potential for image captioning generation to revolutionize content accessibility, understanding, and engagement remains significant, making it an area of ongoing exploration and development.

10. FUTURE SCOPE

In the future, several enhancements can be made to further improve image captioning generation.

Here are some potential areas of focus:

1. **Improved Accuracy and Understanding:** Research can be directed towards developing models that can generate more accurate and contextually meaningful captions for complex images. Advancements in deep learning architectures, such as incorporating attention mechanisms or transformer-based models, can enhance the understanding of visual context and improve caption quality.
2. **Reducing Bias and Increasing Fairness:** Efforts can be made to address biases in image captioning models. Research can focus on developing methods to mitigate gender, racial, cultural, or other biases that may be present in the training data, ensuring fairness and inclusivity in the generated captions.
3. **Enhanced Contextual Understanding:** Future developments can aim to improve the ability of image captioning models to understand and incorporate broader contextual information. This includes considering temporal aspects, relationships between multiple images, and understanding the narrative or story conveyed by a sequence of images.
4. **Fine-Grained and Creative Captions:** Advancements can be made to generate captions that are more nuanced, creative, and expressive. Encouraging models to generate captions with diverse styles, incorporating humor, emotion, or storytelling elements can make the captions more engaging and human-like.
5. **Adaptive and Personalized Caption Generation:** Tailoring image captions based on user preferences, context, or domain-specific knowledge can lead to more personalized and relevant captions. Models can be developed to learn from user feedback and adapt the caption generation process to individual needs and preferences.
6. **Multilingual Image Captioning:** Extending image captioning models to generate captions in multiple languages can broaden their applicability and make them accessible to a more diverse user base. This involves training models on multilingual datasets and developing techniques to handle language-specific nuances and variations.

7. Image Captioning in Low-Resource Settings: Research can focus on developing image captioning models that perform well in low-resource settings, where training data may be scarce. Techniques such as transfer learning, domain adaptation, or leveraging multimodal pretraining can help improve performance in such scenarios.
8. Ethical Considerations and User Control: Future enhancements should prioritize incorporating ethical considerations, giving users control over generated captions, and enabling transparency in the captioning process. User interfaces can be designed to allow users to edit or influence the generated captions to ensure accuracy, fairness, and alignment with individual preferences.
9. Integration with Augmented Reality (AR) and Virtual Reality (VR): Image captioning can be integrated with AR and VR technologies to provide real-time or immersive captioning experiences. This can enhance accessibility, gaming, training, and various other applications in these domains.
10. Cross-Modal Understanding: Advancements in research can focus on developing models that have a better understanding of the relationship between visual and textual modalities. This includes exploring methods for generating captions that go beyond mere descriptions and demonstrate a deeper understanding of visual content, incorporating reasoning, inference, and context. Continued research and development in these areas can pave the way for significant improvements in image captioning generation, making it more accurate, versatile, and capable of meeting the evolving needs of users across various domains and applications.

APPENDIX

A. Source Code

App.py

```
File Edit Selection View Go Run ... ImageCaptionGeneration
app.py 6, M X caption.html U index.html M # style.css U JS script.js M
app.py > ...
1 from pickle import load
2 from numpy import argmax
3 from tensorflow.keras.preprocessing.sequence import pad_sequences
4 from tensorflow.keras.applications.vgg16 import VGG16
5 from tensorflow.keras.preprocessing.image import load_img, img_to_array
6 from tensorflow.keras.applications.vgg16 import preprocess_input
7 from tensorflow.keras.models import Model
8 from tensorflow.keras.models import load_model
9 import os
10 from flask import Flask, render_template, request
11 from werkzeug.utils import secure_filename
12
13 app = Flask(__name__)
14
15 @app.route('/')
16 def home():
17     return render_template("index.html")
18
19 @app.route('/Prediction')
20 def Prediction():
21     return render_template('caption.html')
22
23 @app.route('/PredictCaption', methods=['GET', 'POST'])
24 def upload():
25     if request.method == 'POST':
26         f = request.files['image']
27         basepath = os.path.dirname(__file__)
28         filepath = os.path.join(basepath, "uploads", secure_filename(f.filename))
29         f.save(filepath)
30         text = modelpredict(filepath)
31         return text
32
33 def extract_features(filename):
34     model = VGG16()
```

```
File Edit Selection View Go Run ... ImageCaptionGeneration
app.py 6, M X caption.html U index.html M # style.css U JS script.js M
app.py > ...
44 def word_for_id(integer, tokenizer):
45     for word, index in tokenizer.word_index.items():
46         if index == integer:
47             return word
48     return None
49
50 def generate_desc(model, tokenizer, photo, max_length):
51     in_text = 'startseq'
52     for _ in range(max_length):
53         sequence = tokenizer.texts_to_sequences([in_text])[0]
54         sequence = pad_sequences([sequence], maxlen=max_length)
55         yhat = model.predict([photo, sequence], verbose=0)
56         yhat = argmax(yhat)
57         word = word_for_id(yhat, tokenizer)
58         if word is None:
59             break
60         # in_text += ' ' + word
61         if word == 'endseq':
62             break
63         in_text += ' ' + word
64     # Remove 'startseq' from the beginning and 'endseq' from the end
65     in_text = in_text.replace('startseq', '').replace('endseq', '').strip()
66     return in_text
67
68 def modelpredict(filepath):
69     tokenizer = load(open("tokenizer.pkl", 'rb'))
70     max_length = 34
71     model = load_model("caption.h5")
72     photo = extract_features(filepath)
73     description = generate_desc(model, tokenizer, photo, max_length)
74     return description
75
76 if __name__ == "__main__":
77     app.run(debug=True)
```

Index.html-

```
File Edit Selection View Go Run ... ImageCaptionGeneration
app.py 6,M caption.html U index.html M X # style.css U JS scripts.js M
templates > index.html > html > body > div.container > form > div.center > div.dropzone > input#imageupload-input
1 <!DOCTYPE html>
2 <html>
3
4 <head>
5 <title>Image Upload Form</title>
6 <style>
7   .container {
8     border-radius: 5px;
9     background-color: #f2f2f2;
10    padding: 20px;
11    position: absolute;
12    top: 50%;
13    left: 50%;
14    width: 400px;
15    height: 400px;
16    margin-top: -200px;
17    margin-left: -200px;
18    border-radius: 2px;
19    box-shadow: 4px 8px 16px 0 rgba(0, 0, 0, 0.1);
20    overflow: hidden;
21    background: linear-gradient(to top right, darkmagenta 0%, hotpink 100%);
22    color: #333;
23    font-family: "Open Sans", Helvetica, sans-serif;
24  }
25
26  .center {
27    position: absolute;
28    top: 50%;
29    left: 50%;
30    transform: translate(-50%, -50%);
31    width: 300px;
32    height: 260px;
33    border-radius: 3px;
34    box-shadow: 8px 10px 15px 0 rgba(0, 0, 0, 0.2);
```

```
File Edit Selection View Go Run ... ImageCaptionGeneration
app.py 6,M caption.html U index.html M X # style.css U JS scripts.js M
templates > index.html > html > body > div.container > form > div.center > input.btn
34 box-shadow: 8px 10px 15px 0 rgba(0, 0, 0, 0.2);
35 background: #fff;
36 display: flex;
37 align-items: center;
38 justify-content: space-around;
39 flex-direction: column;
40
41
42 input[type=file],
43 select {
44   width: 100%;
45   padding: 12px;
46   border: 1px solid #ccc;
47   border-radius: 4px;
48   resize: vertical;
49   margin-bottom: 5px;
50 }
51
52 .title {
53   width: 100%;
54   height: 50px;
55   border-bottom: 1px solid #999;
56   text-align: center;
57 }
58
59 h1 {
60   font-size: 16px;
61   font-weight: 300;
62   color: #666;
63 }
64
65 .dropzone {
66   width: 100px;
67   height: 80px;
```

```
File Edit Selection View Go Run ... ImageCaptionGeneration
app.py 6, M caption.html U index.html M X # style.css U JS script.js M
templates > index.html > html > body > div.container > form > div.center > input.btn
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
border: 1px dashed #999;
border-radius: 3px;
text-align: center;
}

.upload-icon {
margin: 25px 2px 2px 2px;
}

.upload-input {
position: relative;
top: -62px;
left: 0;
width: 100%;
height: 100%;
opacity: 0;
}

.btn {
display: block;
width: 140px;
height: 40px;
background: #darkmagenta;
color: #fff;
border-radius: 3px;
border: 0;
box-shadow: 0 3px 0 #hotpink;
transition: all 0.3s ease-in-out;
font-size: 14px;
}

.btn:hover {
background: #rebeccapurple;
box-shadow: 0 3px 0 #deeppink;
}
```

```
File Edit Selection View Go Run ... ImageCaptionGeneration
app.py 6, M caption.html U index.html M X # style.css U JS script.js M
templates > index.html > html > body > div.container > form > div.center > input.btn
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
```

