CSCE 5290: Natural Language Processing
Project Proposal

Image Captioning and Speech Recognition using NLP

Group Members

Ali shah - 11518121
Nithish Kumar Boggula -11559328
Aditya Kapilavai - 11594174

GitHub Repository Link:- https://github.com/Nithishkumar-11/NLP_project

GOALS AND OBJECTIVES:

Motivation:

The project automatically produce a Natural language description of an image when given as input and also To get the image as an output when an input text or voice is given. It has numerous potential uses, including helping people with visual impairments to comprehend images, enhancing the effectiveness of image search engines, and automatically generating descriptions for images in social media and news articles. This technology has the capacity to transform the way we engage with visual content, making it more widely accessible.

This strategy could lower crime and/or accidents since it simulates the human ability to describe visuals using computer language. Because every image might be transformed into a caption before being searched on, automatic captioning could also contribute to the improvement of Google Image Search. Speech recognition is used for converting speech into text, searching for keywords, and finding appropriate pictures using NLP tools. The goal is to create technologies that allow machines to comprehend and produce human language, which could assist in enhancing the interaction between humans and machines.

Objective:

The primary objective of the project "Image captioning and speech recognition using Natural language processing" is to develop a system that can automatically generate a textual description of an image using natural language processing techniques and by using speech recognition which converts the input voice of a person into a text and then based on the text find

the images best suited for that particular description. The system takes an image as input and produces a description that accurately describes the content of the image and vice-versa.
Some specific objectives of this project include:

1. Developing an image recognition system that can identify the objects, people, and other elements present in the image.
2. Developing a speech recognition system to take the voice of a person as input and recognize the input and convert it into text.
3. Creating a natural language processing model that can generate coherent and descriptive captions based on the image features.
4. Integrating the speech recognition module with the text classifier and searching for the keywords.
5. Based on that if an image exactly matches the description then an output is displayed.
6. Integrating the image recognition and natural language processing models to create a unified system that can generate captions for any given image.
7. Evaluating the performance of the system through various metrics such as accuracy, fluency, and relevance of the generated captions.
8. Exploring ways to improve the system's performance, such as incorporating attention mechanisms, using more advanced natural language processing techniques, or incorporating user feedback to refine the captions.

Significance:

This project has several significant implications and potential benefits. Some of the key significances of the project are:

- ➔ **Accessibility**: This project's primary advantage is its ability to improve the accessibility of visual content for individuals with visual impairments. The automatic generation of text descriptions for images and speech-to-text conversion can significantly enhance inclusivity and accessibility of visual content.
- ➔ **Content Discovery**: One potential application of this technology is to facilitate the organization and navigation of extensive image collections, thereby simplifying the process of locating and browsing visual content. This could prove particularly

advantageous in fields such as art, fashion, and photography, which rely heavily on images as a means of communication.

- ➔ **Automation**: Additionally, this technology could be utilized to automate the creation of captions for large quantities of images, resulting in substantial savings in terms of time and resources. This would be particularly beneficial in industries like advertising and e-commerce, where numerous images are employed to advertise and showcase various products and services.
- ➔ **Personalization**: Image captioning can be used to generate captions that are tailored to the preferences and interests of individual users. This can help create more personalized user experiences and improve engagement with visual content.
- ➔ **Research**: Image captioning can be used to advance research in computer vision, natural language processing, and machine learning. The development of accurate and efficient image captioning systems can help advance the state-of-the-art in these fields and lead to new breakthroughs and innovations.

Overall, the project has the potential to make significant contributions to various fields and improve the accessibility, usability, and relevance of visual content.

Features:

The features can vary depending on the specific implementation and approach used, but here are some common features that are often included in such projects:

Image recognition: To identify the various components within an image, the system employs an image recognition model. This process commonly involves utilizing convolutional neural networks (CNNs) to extract distinctive characteristics from the image and then implementing machine learning approaches to categorize the objects, individuals, and other relevant elements that appear within the image.

Speech recognition for converting speech to text: Speech recognition uses machine learning algorithms and neural networks to recognize and transcribe spoken words into text. This technology has several applications, including voice assistants, and speech-to-text transcription services. Some popular speech recognition APIs include Google Cloud Speech-to-Text, Amazon Transcribe, and Microsoft Azure Speech Services.

Natural language processing: The system uses natural language processing techniques to generate a textual description of the image. This typically involves using recurrent neural networks (RNNs) or transformer models to generate a sequence of words that describe the objects in the image.

Searching for keywords in text: Once we have converted speech to text, you can use natural language processing (NLP) tools to search for keywords. One popular NLP tool used for keyword extraction is the Python Natural Language Toolkit (NLTK). The NLTK includes a range of algorithms for processing natural language text, including tools for tokenization, part-of-speech tagging, and named entity recognition. These tools can be used to identify and extract important keywords from the text.

Attention mechanisms: To ensure that the generated caption accurately reflects the contents of the image, the system may use attention mechanisms that focus on different parts of the image when generating different parts of the caption.

Fine-tuning: The system may be fine-tuned using transfer learning techniques, which involves pre-training on large datasets and then fine-tuning on a smaller dataset to improve performance on a specific task.

Evaluation metrics: To evaluate the performance of the system, various metrics are used, such as BLEU score, METEOR score, ROUGE score, and CIDEr score, which measure the similarity between the generated caption and a reference caption.

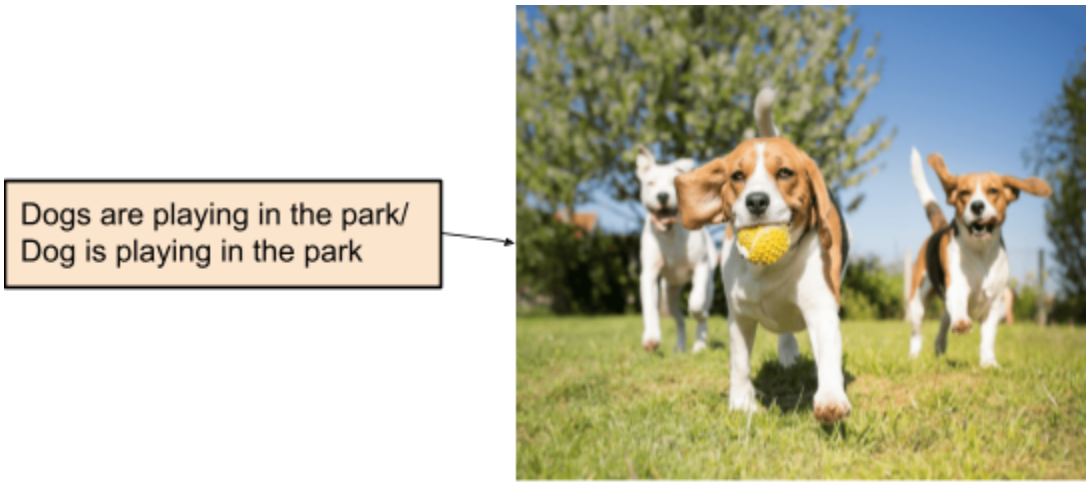
User interface: Depending on the intended use case, the system may have a user interface that allows users to upload images and view the generated captions. The user interface may also allow users to provide feedback on the quality of the generated captions, which can be used to improve the system.

Finding appropriate pictures: To find appropriate pictures based on the keywords, we can use image search APIs such as Google Image Search or Bing Image Search. These APIs allow you to programmatically search for images based on specific keywords or phrases. We can also use computer vision APIs such as Microsoft Azure Cognitive Services or Google Cloud Vision to analyze images and identify relevant features, such as colors, shapes, and objects.

On the whole, the project involves the integration of various computer vision and natural language processing techniques to generate accurate and descriptive captions for images. The features of the system will depend on the specific implementation and requirements of the use case.

Visualization:

In the above fig, If the image input is given, Caption will be generated.



Text to image based on the keywords

If the input words/sentences are given, images/related images will the output.

a little girl in a pink dress going into a wooden cabin .
a little girl climbing the stairs to her playhouse .
a little girl climbing into a wooden playhouse .
a girl going into a wooden building .
a child in a pink dress is climbing up a set of stairs in an entry way .

two dogs on pavement moving toward each other .
two dogs of different breeds looking at each other on the road .
a black dog and a white dog with brown spots are staring at each other in the street .
a black dog and a tri-colored dog playing with each other on the road .
a black dog and a spotted dog are fighting

young girl with pigtails painting outside in the grass .
there is a girl with pigtails sitting in front of a rainbow painting .
a small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
a little girl is sitting in front of a large painted rainbow .
a little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .

man laying on bench holding leash of dog sitting on ground
a shirtless man lies on a park bench with his dog .
a man sleeping on a bench outside with a white and black dog sitting next to him .
a man lays on the bench to which a white dog is also tied .
a man lays on a bench while his dog sits by him .

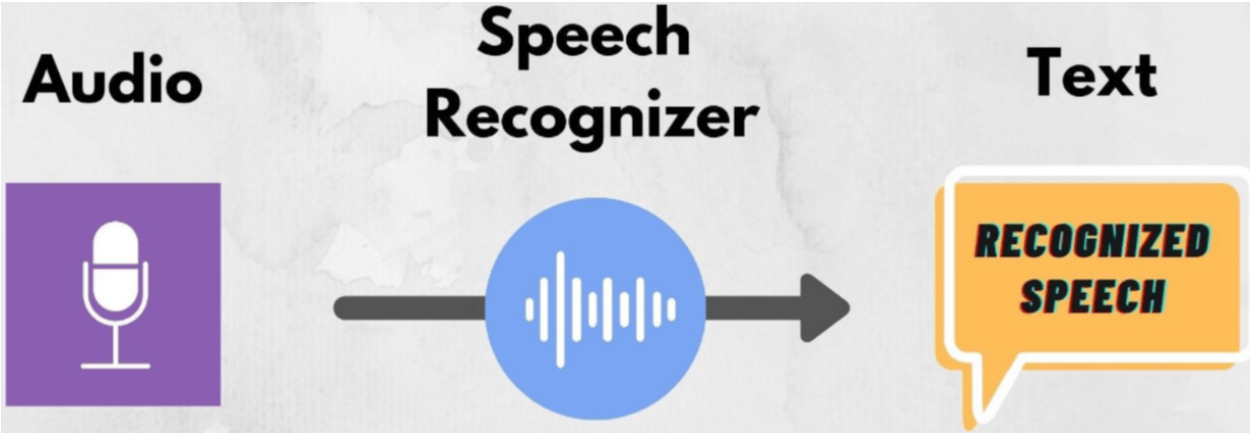
the man with pierced ears is wearing glasses and an orange hat .
a man with glasses is wearing a beer can crocheted hat .
a man with gauges and glasses is wearing a blitz hat .
a man wears an orange hat and glasses .
a man in an orange hat starring at something .

The above fig demonstrates the dataset of input and training Captions.

NLP tools for speech recognition and keyword extraction:

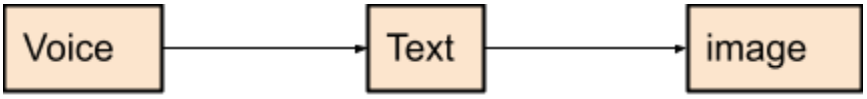
Some popular NLP tools for speech recognition and keyword extraction include:

- Google Cloud Speech-to-Text API: It is a cloud-based speech recognition API that transforms audio into text.
- IBM Watson Speech to Text: A speech-to-text API which supports multiple languages and has built-in custom language models.
- Spacy: A Python library for NLP tasks, including tokenization, part-of-speech tagging, and named entity recognition.
- Gensim: A Python library for topic modeling, similarity detection, and keyword extraction from text.
- NLTK: A Python library for NLP tasks, including tokenization, part-of-speech tagging, named entity recognition, and sentiment analysis.



This can be applied to both Text to speech and speech to text respectively. Once the speech or audio has been given as text then the text is tokenized and we search for keywords in the text. Based on the keywords we can match the text with the image.

Inorder to make it work properly the model needs to be trained and must recognize the keywords and speech must be exactly identified. There can be many features added to this project in the coming future.



Speech Recognition

References:

Enhancing Descriptive Image Captioning with Natural Language Inference https://aclanthology.org/2021.acl-short.36/

Generating Image Captions based on Deep Learning and Natural language Processing https://ieeexplore.ieee.org/abstract/document/9596486

Text to Image Synthesis for Improved Image Captioning https://ieeexplore.ieee.org/abstract/document/9416431

An Efficient Text-Based Image Retrieval Using Natural Language Processing (NLP) Techniques https://link.springer.com/chapter/10.1007/978-981-15-5400-1_52

D.V.T and V.R, "Retrieval Of Complex Images Using Visual Saliency Guided Cognitive Classification", J. Innov. Image Process, vol. 2, no. 2, pp. 102-109, Jun. 2020. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://irojournals.com/iroiip/V2/I2/05.pdf