Image Captioning and Speech Recognition using NLP

Group Members

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

GitHub Repository Link:- https://nithishkumar-11.github.io/NLP_project/

Motivation:

A key objective of this project is to automatically produce a natural language description of an image and To get the image as an output when an input text or voice is given . The main goal of this project is to develop a method that implements natural language inference to motivate models to generate more comprehensive captions. 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. This even works for the voice taken as input and it gives us the output as images when they are given as input. The ultimate aim is to recognize what the picture says based on its characteristics and then when a voice or text is given let's try to visualize that into an image. Speech recognition is used for converting speech into text, searching for keywords, and finding appropriate pictures using NLP tools.

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 converting the 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.

Overall, the project aims to develop a practical and effective system for image captioning and speech recognition that can be used in a variety of applications, such as aiding the visually impaired or assisting with image indexing and retrieval in large datasets.

For humans, It is really easy when an image is given rather than a text or paragraph. So, our goal is to make things easier and simpler. Nowadays we want things to become automated as this model is perfectly suited for people who are facing challenges hearing things and have trouble explaining things.

It helps people to get an image and even get something from the image. This is a world of innovation and NLP plays a huge role in the advancement of technology.

We need a Mic or input device to take the input as voice and once the input is given the speech needs to be recognized by using the speech recognition module and then it is parsed as a text.

Significance:

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

➔ **Accessibility**: One of the most significant benefits of this project is its potential to make visual content more accessible to people with visual impairments. By automatically generating textual descriptions of images and speech to text, this project can help make visual content more inclusive and accessible.

➔ **Content Discovery**: This can be used to index and organize large collections of images, making it easier for people to discover and navigate through visual content. This can be especially useful in fields such as art, fashion, and photography,etc… where images are the primary source of communication.

➔ **Automation**: It can be used to automate the process of generating captions for large volumes of images, which can save significant amounts of time and resources. This can be especially useful in industries such as advertising and e-commerce, where a large number of images are used to showcase 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: The system uses an image recognition model to identify the objects, people, and other elements present in the image. This typically involves using convolutional neural networks (CNNs) to extract features from the image and then applying machine learning techniques to classify the objects.

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.

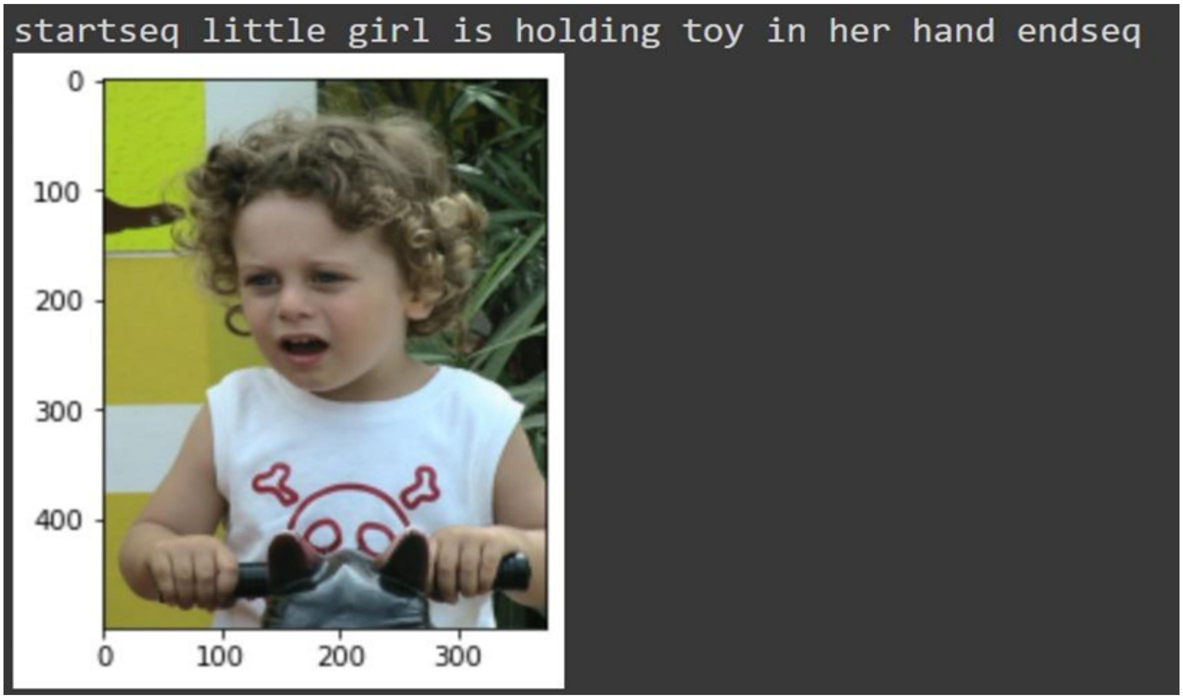
Evaluation metrics: To evaluate the performance of the system, various metrics are used, such as BLEU score, METEOR score, ROUGE score, and CIDEr score, which measure the similarity between the generated caption and a reference caption.

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.

Overall, 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.

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.

Visualization:



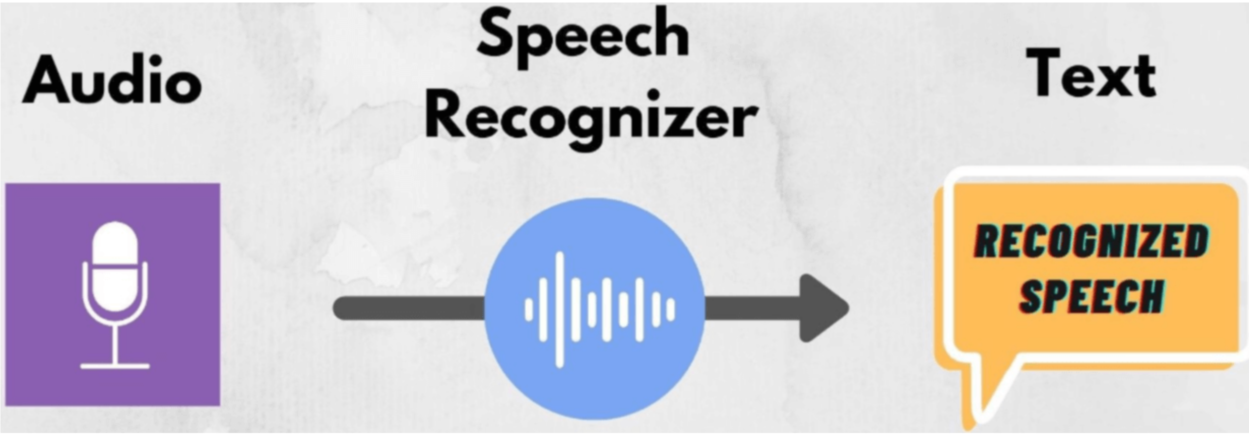
```
startseq little girl is holding toy in her hand endseq
```

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.

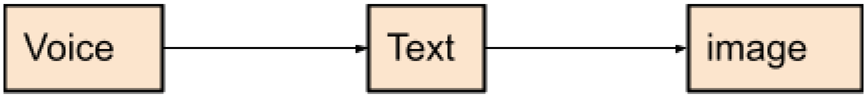
- NLTK: A Python library for NLP tasks, including tokenization, part-of-speech tagging, named entity recognition, and sentiment analysis.



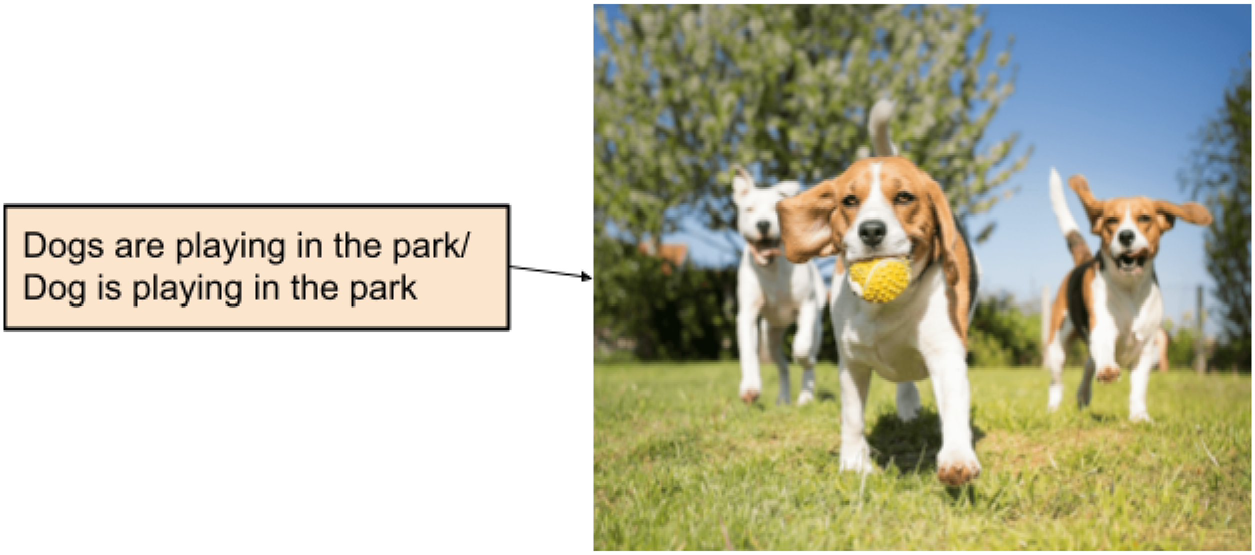
This can be applied to both Text to speech and speech to text respectively. Once the speech or audio has been given as text then the text is tokenized and we search for keywords in the text. Based on the keywords we can match the text with the image.

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



Dogs are playing in the park/
Dog is playing in the park

Text to image based on the keywords