Tunisian Republic
Ministry of Higher Education and Scientific Research
Tunis El Manar University
National Engineering School of Tunis

Spatial Computing Project

Spatial image captioning using large language models (LLM)

Elaborated by :

Wajih OUNIS

Elbokhary Mouhamed AHMEDOU

3rd Year Computer Science

Scholar Year: 2023/2024

Contents

List of Figures

Acknowledgement

We extend our heartfelt gratitude to Dr. Mohsen Ali FRIHIDA for covering the subject of spatial computing and introducing me and our peers to such incredible field and impeccable application of computer science and artificial intelligence in the spatial field.

Abstract

With the explosive evolution in large language model it has become possible to describe spectral images using these llms. In fact, an llm is capable with a little fine-tuning to provide a comprehensive description for the image in question.

Introduction

In this report of the project which is around captioning spectral images using large language models we will go through the steps of designing and implementing this project. This report is divided into multiple sections which are the model acquiring, spectral image processing, fine-tuning the model and serving the application through an api.

Chapter 1

Prototype of the project

1.1 Introduction

In order to get a first minimal piece of functionning software, we have developped a prototype for the project which is a simple web interface that is used to caption rgb images.

1.2 Defitintions

In this section of the prototype we will go through definitions of the terms and technologies used in this project.

1.2.1 Large Language Model

A large language model (LLM) is a language model notable for its ability to achieve general-purpose language understanding and generation. LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process. LLMs are artificial neural networks following a transformer architecture. [1]

1.2.2 Spectral images

Spectral imaging is imaging that uses multiple bands across the electromagnetic spectrum. While an ordinary camera captures light across three wavelength bands in the visible spectrum, red, green, and blue (RGB), spectral imaging encompasses a wide variety of techniques that go beyond RGB. Spectral imaging may use the infrared, the visible spectrum, the ultraviolet, x-rays, or some combination of the above. It may include the acquisition of image data in visible and non-visible bands simultaneously, illumination from outside the visible range, or the use of optical filters to capture a specific spectral range. It is also possible to capture hundreds of wavelength bands for each pixel in an image.[2]

1.3 Vision Models

Vision models are a segment of large language models that are trained on visual data and are capable of performing visual data tasks. Some of these models are open source like gpt2 however some of them require payments for the api key. The model used in this project is a gpt2 based model which loaded through the hugging face transformers library a VisionEncoderDecoderModel. The model is built in a class so it could be later on improved and easily imported and used with other modules like the api serving module.

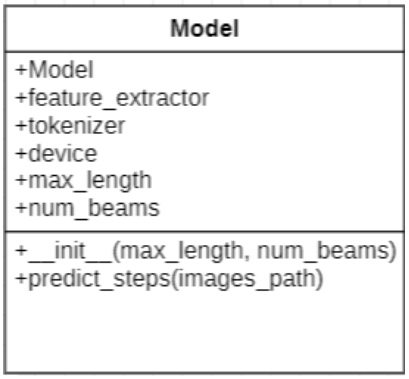

Figure 1.1: Model class

1.4 API class

As this application is still in prototyping phase, the pick for building the api had to be a light weight, straight forward framework so the choice was to go with flask. Flask is a lightweight, easy to use and simple python framework for backend development which has recently became html tolerant thus it can be used to develop both the user interface and the backend of the website. The main function of the api code which is shown in the figure below, designated under the '/upload' route and utilizing the Flask framework, facilitates the seamless upload and processing of files. Users can submit files, and the program ensures the presence and non-emptiness of the uploaded file before proceeding.

Upon successful validation, the program securely saves the file in a predefined upload folder. Notably, it incorporates a file type check using the 'allowed_file' function to ensure compatibility with the application.

An additional feature of this program is its integration with a language modeling component, specifically the 'LLM' (Language Learning Model). After file processing, the program leverages LLM to generate captions, providing users with textual insights related to the uploaded content.

The response to the user includes rendering an HTML template ('index.html') with the generated caption ('caption'). This report outlines the functional aspects of the program, emphasizing its file handling capabilities and the integration of linguistic analysis for enhanced user interaction.

```python
@app.route('/upload', methods=['POST'])
def upload_file():
    if 'file' not in request.files:
        return redirect(request.url)

    file = request.files['file']

    if file.filename == '':
        return redirect(request.url)

    if file and allowed_file(file.filename):
        file_path = os.path.join(app.config['UPLOAD_FOLDER'], file.filename)
        file.save(file_path)


        caption = LLM.predict_step([file_path])

        return render_template('index.html', filename=file_path, caption=caption)

    return redirect(request.url)
```

Figure 1.2: Main API function

1.5 Image processing class

This image processing class is still at an early phase of development and not yet usable as intented, however this class is meant to take care of spectral image processing before passing them to the model.

Figure 1.3: Image process class

1.6 Conclusion

In conclusion, as a prototype this end to end application is performing well. And its object oriented SOLID structure opens it to more extension and functionalities in the next versions of the application.

Conclusion

In conclusion, this project signifies a significant advancement in the realms of image captioning as the revolution of large language has opened a large spectrum of possibilities and broad horizon of new opportunities and applications whether in scientific research, business or literature. The challenge of the this artificial intelligence explosive rise has deviated from being focused around technological advancement to the potential immoral use of this "dangerous weapon" and how to mitigate that.

Bibliography

[1] wikipidea.

[2] wikipidea.