ETH zürich

Institute of Cartography
and Geoinformation

Automatic Map Storytelling with Generative Pre-trained Transformer (GPT) Models

Group members: Ziyi Liu, Claudio Affolter
Leading Professor: Prof. Dr. Lorenz Hurni
Advisors: Sidi Wu, Dr. Yizi Chen

1 Introduction

Maps serve as valuable tools for acquiring information and knowledge about the world. However, the diversity of map types as well as the wide range of map styles and thematic/temporal contexts make it challenging for non-experts to easily identify and understand maps, particularly historical ones.
Fortunately, recent advances in image captioning methods such as CLIP (Contrastive Language-Image Pre-Training) **[1]** and ClipCap (CLIP Prefix for Image Captioning) **[2]** are promising to help overcome these challenges. These methods combined with GPT models pre-trained on huge datasets can automatically generate image captions.
Nevertheless, for historical maps, the generated captions are either too simple, too general, or even wrong.

2 Goals

The goal of this project is to explore GPT models for map storytelling, where given a historical input map, the model should generate a caption answering the following questions:

1. **Where?** The area which is depicted on the input map.
2. **What?** The map type as well as the topic or style.
3. **When?*** The century in which the map was created.
4. **Why?** The purpose of the map.

The focus was set on topographic maps created between the 16th and 19th century, and pictorial maps featuring different topics.
* Only for topographic maps.

3 Method Overview

- Assembling over 4000 historical maps using David Rumsey Map collection **[3]** and preparing datasets with manually structured captions.
- Fine-tuning CLIP on a maximum of around 4000 maps and training ClipCap on a subset of around 750 maps.
- Evaluating performance by measuring accuracy for the CLIP model and assessing cosine similarity between the generated and target captions for ClipCap.
- Utilizing the assembled map collection to develop distinct models for each of the specific questions outlined in the *Goals* section.
- Creating decision tree structure for prediction in order to construct a complete caption with the help of GPT-3.5 **[4]** from keywords predicted by the fine-tuned CLIP models.

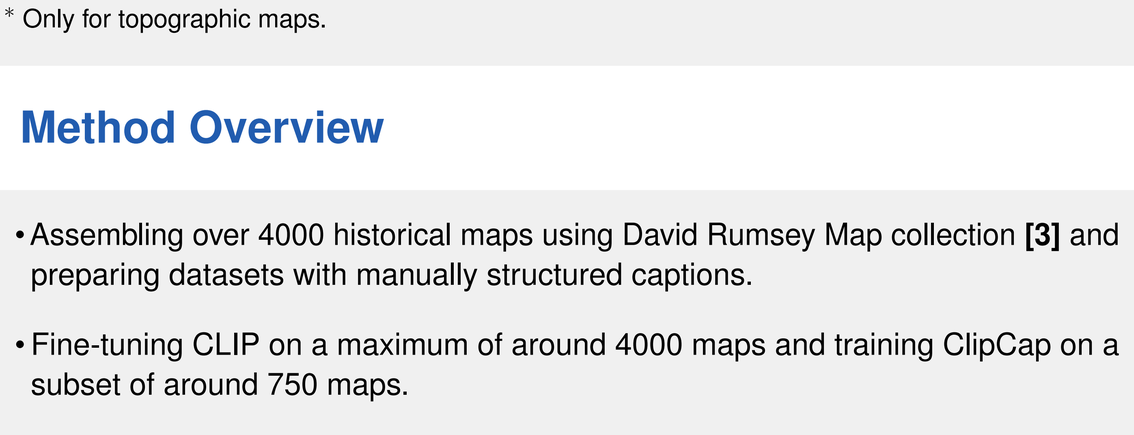

Figure 1: Summary of approach used in CLIP

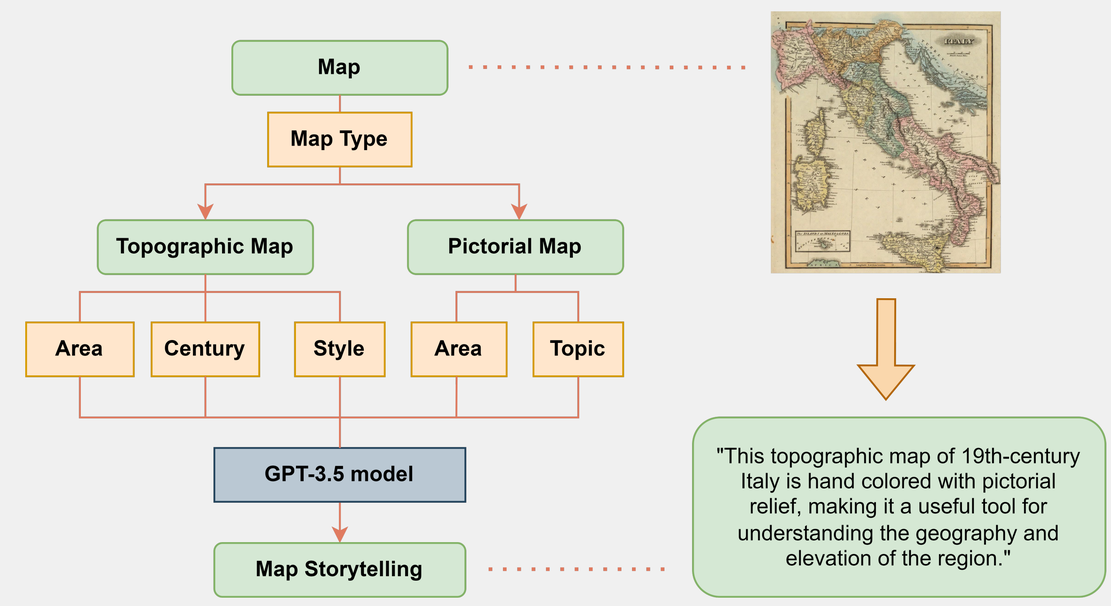

Figure 2: Overview of decision tree structure.
Input is a historical map and output is a caption answering *Where?, What?, When?* and *Why?*.

4 Results and Discussion

Category	Base CLIP	Fine-tuned CLIP
Map Type *(2)*	0.39	**0.97**
Area, *Topo. (27)*	0.32	**0.78**
Style, *Topo. (6)*	0.35	**0.86**
Century, *Topo. (4)*	0.37	**0.78**
Area, *Pic. (2)*	**0.97**	0.94
Topic, *Pic. (13)*	0.39	**0.76**

Table 1: Comparison of prediction accuracy achieved (per category) with base CLIP model and fine-tuned CLIP models. Topo. = Topographic map, Pic. = Pictorial map. The numbers in brackets next to the categories display the amount of classes within a category e.g., the topographic maps depict 27 different areas.

This quantitative comparison shows that the fine-tuned CLIP models clearly outperform the base CLIP model in five out of six categories.

5 Conclusion

Developing a method for automatic map storytelling using only ClipCap did not lead to satisfying results (which ultimately led to this model being discarded), as ClipCap's underlying image encoder CLIP model was pre-trained on the COCO image dataset **[5]**. This dataset consists entirely of images featuring real life situations that are too distinct from historical maps. The final developed image captioning method utilizes a decision tree prediction structure consisting of individual CLIP models and the generative ability of GPT-3.5. This approach is capable of generating a descriptive caption answering the four questions introduced in the *Goals* section, thereby facilitating detailed storytelling.

References

[1] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[2] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning, 2021.

[3] Rumsey, David, and Cartography Associates. David Rumsey Map Collection.

[4] OpenAI. GPT-3.5, 2022.

[5] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context, 2014.