Automatic Illustration of Text via Multimodal Interaction

Dominykas Meistas - 2404288

December 17, 2021

1 Status report

1.1 Proposal

1.1.1 Motivation

Illustrating a concept with an image is an effective way to increase one's understanding of that field. Humans can perceive images a lot faster than long paragraphs of text, which makes the idea of this project very valuable. Our tool could be helpful in many different scenarios. It could help a student make a good slideshow presentation by offering him a choice of relevant images to include. It would make the presentation more understandable and enjoyable and allow the student to save much time by not needing to search for images by himself. It could also help anyone writing any document or paper that could be enhanced by using images.

1.1.2 Aims

The project aims to create an interface that would allow users to write text. It would then extract the keywords from that text section, acquire relevant images based on those keywords, and suggest them to the user, which he could include in his work. The goal is that the extracted keywords accurately represent the text section, and the retrieved images match the keywords and the context of the paragraph. The model must also know the similarities between pairs of words and images. Hence it should be able to retrieve relevant images even if a specific keyword has not occurred in the training set of the model.

1.2 Progress

- Background research conducted on multimodal interaction, multimodal systems.
- Dataset to train, test, evaluate the model chosen: Wikipedia-based Image Text (WIT) dataset will be used.
- Language to implement the model chosen: the project will be implemented in Java.
- Text index chosen: Apache Lucene will be used to create a text index and retrieve relevant documents. However, Terrier might be used later because it supports the neural ranking models, which will be necessary to implement text-text and text-image semantic understanding.
- Initial text index pipeline implemented: given a Wikipedia text section, the pipeline returns the images from the Wikipedia documents that are most relevant to the text section given.

1.3 Problems and risks

1.3.1 Problems

- The biggest problem was understanding how the multimodal space works and deciding what approach we should take to build the model. We dealt with this problem by reading many research papers, analysing them.
- Another problem occurred when building the Lucene text index. The program did not run correctly because our Maven dependencies were not updated, which caused some delay in the project development.

1.3.2 Risks

- Might be challenging to implement text-text and text-image semantic similarities and create a multimodal space, as it is a difficult concept to grasp and might require a lot of time and effort. **Mitigation:** will start working on this issue right away to have as much time to complete this task.
- Besides our primary method, we might need to find an alternative way to assess our model. The current method we have in mind can only evaluate if the returned image is the exact image in the text section on the Wikipedia page. However, it does not know if other retrieved images are relevant or not. **Mitigation:** do some extra reading about this issue, talk with the supervisor, and come up with a new idea to evaluate the model.

1.4 Plan

Semester 2

- Week 1-4: build the multimodal semantic space.
  - **Deliverable:** a model with a text-text and text-image semantic understanding, meaning that it understands which words have a similar meaning and which images represent them.
- Week 5-6: evaluate the product.
  - **Deliverable:** different plots, graphs, tables, other types of visualisation, that would show how good our product performs compared to our expectations and other products on the market.
- Week 7-8: implement any extra features and finalise the product.
  - **Deliverable:** the final product with all required features.
- Week 9-11: write the dissertation.
  - **Deliverable:** a final dissertation.