Project Proposal:
A- Composed Image Retrieval

Aymane El Firdoussi
Student

Lucas Ventura
Supervisor

Gül Varol
Supervisor

Abstract

Composed Image Retrieval (CoIR) has recently gained much attention by the computer vision community. It involves retrieving images based on a complex multi-type query comprised of a reference image and a text-based description or modification of this latter. CoIR is inherently challenging as it requires using advanced techniques to learn and integrate both visual and textual information.

1 Plan of the project

The primary objective of this project is to develop a method that enhances the performance of current state-of-the-art CoIR models on the CIRR dataset, which will serve as our benchmark metric. To achieve this, we will explore various research directions. The project plan can be outlined as follows:

1. Examining related work done in this problem and reading the reference papers, especially Ventura et al. (2024) and Liu et al. (2021), in order to get a better understanding of the current challenges of CoIR.
2. Creating our baseline results by reproducing the experiment described in Table 3 in (Ventura et al. (2024)), using smaller batch size and lower compute power (training on a single GPU). And to do so, we will rely on the CoVR codebase[1] provided by the phD student **Lucas Ventura**.
3. Investigating on the impact of the image `q`, text `t` and query `f(q, t)` embeddings on the performance of the model through:
   - Learning (tentative) a combination of all the three embeddings that leads to better performance of the model (instead of relying solely on the query `f(q, t)`).
   - Experimenting with different pooling techniques to aggregate the learnable queries embeddings, rather than simple averaging. These include max-pooling or attention-based pooling.
4. Reporting the obtained results and, based on these findings, suggesting new research directions that could further enhance the model's performance and that could be tested in a future work.

References

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[1] https://github.com/lucas-ventura/CoVR