Title: Multimodal Summarization

Author: Mateusz Krubiński

Department: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Pavel Pecina, Ph.D.,
Institute of Formal and Applied Linguistics

Abstract: The task of Multimodal Summarization aims to fuse disjoint information from several sources (modalities) and distill it into a concise and precise summary. In our research, we approach a text-centric variant that requires textual content in both the input and the output, i.e., the summary. Specifically, we focus on the Multimodal Summarization with Multimodal Output (MSMO) approach, which summarizes a textual document accompanied by either a collection of images or a short video into a textual summary accompanied by a single image. On the modeling side, we are interested in supervised formulations that explore a single neural model to generate the multimodal summary end-to-end, i.e., by simultaneously processing textual and visual modalities. Considering the task's novelty, it still lacks the core components of a well-established field, such as standardized benchmarks (datasets), publicly available baseline models, and even task-specific metrics. Therefore, our main contributions are aimed at performing basic research to establish foundations for future work. Namely, we: i) curate and publish a large-scale video-based dataset for MSMO; ii) perform experiments to establish the role of pre-training and the influence of the (quality of) visual input on the (quality of) textual output; iii) design a human evaluation framework for MSMO, and propose a novel metric for evaluating the quality of textual output; iv) propose a simplified, multi-task formulation of MSMO, that unifies the image-based, video-based, and text-only variants with a single architecture.

Keywords: summarization, text summarization, vision-language modeling, multimodal data