DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, Yale Song

March 3, 2025

Abstract

This paper introduces DOC2PPT, a task for generating presentation slides from scientific documents that involves summarization, image retrieval, and organizing content for presentation. The authors propose a hierarchical sequence-to-sequence approach that outperforms strong baselines, with an accompanying dataset of approximately 6,000 document-slide pairs.

Introduction

Creating presentations requires complex reasoning skills to summarize and visually arrange information, which raises the question of whether machines can replicate this process. The authors elaborate on the unique challenges in automated slide generation from documents, considering both textual and visual elements.

Related Work

The authors provide an overview of relevant literature in vision-and-language modeling, document summarization, and multimodal summarization, highlighting that previous work does not adequately address the unique challenges posed by generating structured outputs, specifically presentation slides. They emphasize the novelty of the DOC2PPT task within this context.

Approach

The DOC2PPT model involves reading a document, paraphrasing content for slides, and strategically determining object placement on each slide. The architecture includes a modular design with components for document reading, progress tracking, object placement, and paraphrasing.

Model

The model employs a Document Reader that encodes the document's text and figures, a Progress Tracker that governs the hierarchical structure, and an Object Placer that determines slide content and layout. The Paraphraser reformulates text to be concise and suitable for presentation.

Training

The authors establish a training methodology that balances structural and content similarity, ensuring that the generated slides closely follow the format and substance of the ground truth. Various components of loss are defined to assess the quality of selected content and layout.

Inference

During inference, a multimodal projection head helps refine the generated slides by adding relevant figures and removing irrelevant ones based on a set threshold. This post-processing step enhances the final slide deck's informativeness.

Dataset

The dataset includes 5,873 pairs of documents and slide decks sourced from various research communities, providing a comprehensive resource for training and evaluation. The authors outline the data collection and extraction processes ensuring robustness and reliability.

Experiments

The authors propose new evaluation metrics specifically for slide generation, performing comprehensive assessments of their approach against various baselines. They also conduct human evaluations to gauge perceived slide quality.

Results and Discussions

The results show that the hierarchical modeling employed in DOC2PPT outperforms flat models across various evaluation metrics. The authors also discuss the effectiveness of individual components in the architecture, such as paraphrasing and layout prediction.

Conclusion

The authors highlight the significance of their work in advancing automatic slide generation, addressing multimodal challenges, and contributing a substantial dataset and evaluation metrics for the community. They anticipate that DOC2PPT will foster improvements in vision-and-language understanding.