

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.

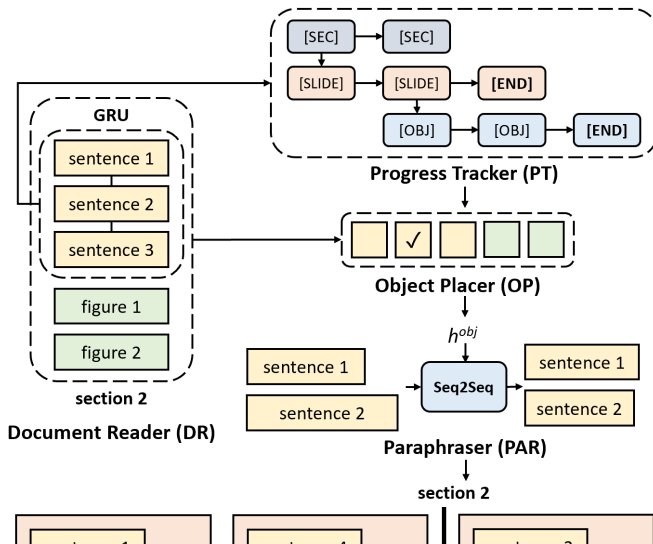
Results and Discussions

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.

Extracted Figure



Extracted Figure

<p>Introduction</p> <ul style="list-style-type: none"> Normalizing flows are a simple approach to generative modeling The latent space is chosen to be Gaussian that is chosen to be $\mathcal{N}(1,1)$ What is latent space, is hard to understand How to infer z (p.17) The density over the decision boundaries should be low 	<p>Introduction</p> <ul style="list-style-type: none"> Normalizing flows are a simple approach to generative modelling The latent space is chosen to be Gaussian that is chosen to be $\mathcal{N}(1,1)$ What is latent space, is hard to understand How to infer z (p.17) The decision over the decision boundary should be low 	<p>Model Analysis</p> <p>Finally, in Section 3.5, we discuss a feature visualization technique that can be used to interpret the features learned by FlowNet.</p> <p>These observations suggest that, in interpreting the model features to part the decision boundaries in the low density region of the data space.</p> <p>One way to analyze the latent representation can be achieved by FlowNet, the evaluate the distributions of the features for the supervised and unsupervised conditions (see Fig. 3.5).</p>	<p>Model Analysis</p> <ul style="list-style-type: none"> A feature visualization technique is used to interpret the features learned by FlowNet. The result aims to find the decision boundaries in the data space. FlowNet: latent representation space learned by FlowNet The results the distributions of the features for the supervised and unsupervised methods. 																				
<p>Methods</p> <ul style="list-style-type: none"> A simple context is related with the user flow function Modify the user flow function with the student network What is the best schedule model performance? Testing the teacher network for different configurations Support new sets of images are given by 	<p>Methods</p> <ul style="list-style-type: none"> A student network is trained with the user flow function Modify the teacher network with the student network What is the best schedule model performance? Testing the teacher network in all feature generated images Support new sets of images are given by 	<p>Conclusion</p> <p>In this work, we have developed an efficient user representation learning method that significantly improves the performance of user image representation on Facebook, simultaneously learning semantic, syntactic, and generative representations.</p> <p>The user representation being generated effectively, between both synthetic and real images and some subrepresentations to improve the performance of the model without the inclusion of additional or irrelevant information.</p> <p>Next, we will extend this proposed through a simple data augmentation (i.e., multi-scale image crops) and apply it to learn performance on user representation tasks.</p> <p>In a result, we will extend the model to an open problem of all three OpenImage benchmarks without the need to fine-tune or any specific design or model task.</p> <p>Acknowledgments We would like to thank the support from Google Mobile Vision platform.</p>	<p>Conclusion</p> <ul style="list-style-type: none"> The proposed network user representation method that significantly improves the performance of user image representation on Facebook. Effectively/efficiently learn synthetic and real images and some subrepresentations. FlowNet (FlowNet) is a generative model. The model achieves the best performance on all three OpenImage benchmarks. Google Mobile Vision and Google Assistant platform. <table border="1"> <thead> <tr> <th></th> <th>FlowNet</th> <th>FlowNet (w/ 100k crop)</th> <th>FlowNet (w/ 100k crop)</th> </tr> </thead> <tbody> <tr> <td>Human ImageNet (100k crop)</td> <td>84.4</td> <td>84.4</td> <td>84.4</td> </tr> <tr> <td>Human ImageNet (100k crop)</td> <td>84.4</td> <td>84.4</td> <td>84.4</td> </tr> <tr> <td>Human ImageNet (100k crop)</td> <td>84.4</td> <td>84.4</td> <td>84.4</td> </tr> <tr> <td>Human ImageNet (100k crop)</td> <td>84.4</td> <td>84.4</td> <td>84.4</td> </tr> </tbody> </table>		FlowNet	FlowNet (w/ 100k crop)	FlowNet (w/ 100k crop)	Human ImageNet (100k crop)	84.4	84.4	84.4	Human ImageNet (100k crop)	84.4	84.4	84.4	Human ImageNet (100k crop)	84.4	84.4	84.4	Human ImageNet (100k crop)	84.4	84.4	84.4
	FlowNet	FlowNet (w/ 100k crop)	FlowNet (w/ 100k crop)																				
Human ImageNet (100k crop)	84.4	84.4	84.4																				
Human ImageNet (100k crop)	84.4	84.4	84.4																				
Human ImageNet (100k crop)	84.4	84.4	84.4																				
Human ImageNet (100k crop)	84.4	84.4	84.4																				

w/o TIM Post Proc.

w/ TIM Post Proc.

w/o PAR

w/ PAR