Visual Storytelling with Question-Answer Plans

Mirella Lapata

University of Edinburgh, Edinburgh

Abstract

Visual storytelling aims to generate compelling narratives from image sequences. Existing models often focus on enhancing the representation of the image sequence, e.g., with external knowledge sources or advanced graph structures. Despite recent progress, the stories are often repetitive, illogical, and lacking in detail. To mitigate these issues, we present a novel framework which integrates visual representations with pretrained language models and planning. Our model translates the image sequence into a visual prefix, a sequence of continuous embeddings which language models can interpret. It also leverages a sequence of question-answer pairs as a blueprint plan for selecting salient visual concepts and determining how they should be assembled into a narrative. Automatic and human evaluation on the VIST benchmark (Huang et al., 2016) demonstrates that blueprint-based models generate stories that are more coherent, interesting, and natural compared to competitive baselines and state-of-the-art systems.

Short Bio

Professor Mirella Lapata is a faculty member in the School of Informatics at the University of Edinburgh. She is affiliated with the Institute for Communicating and Collaborative Systems and the Edinburgh Natural Language Processing Group. Her research centers on computational models for the representation, extraction, and generation of semantic information from structured and unstructured data. This encompasses various modalities, including text, images, video, and large-scale knowledge bases. Prof. Lapata has contributed to diverse applied Natural Language Processing (NLP) tasks, such as semantic parsing, semantic role labeling, discourse coherence, summarization, text simplification, concept-to-text generation, and question answering. Using primarily probabilistic generative models, she has employed computational models to investigate aspects of human cognition, including learning concepts, judging similarity, forming perceptual representations, and learning word meanings. The overarching objective of her research is to empower computers to comprehend requests, execute actions based on them, process and aggregate large datasets, and convey information derived from them. Central to these endeavors are models designed for extracting and representing meaning from natural

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'24 Workshop, Glasgow (United Kingdom), 24-March-2024.

mlap@inf.ed.ac.uk (M. Lapata)
https://homepages.inf.ed.ac.uk/mlap/ (M. Lapata)

language text, internally storing meanings, and leveraging stored meanings to deduce further consequences.