Academic Paper Summarization

Botong Zhang, Yang Dong, Han Yang, Yuxuan (Michael) Gao, Ruixi Gong

Introduction

- **Topic:** Our project delves into automatic text summarization, specifically focusing on summarizing scientific and academic papers.
- **Problem Statement:** The challenge lies in effectively condensing lengthy academic texts into concise, coherent summaries without losing crucial information.
- **Example:** Consider a dense, 20-page scientific paper. Our goal is to generate a brief summary that encapsulates the key findings, methodology, and conclusions, making the content more accessible.
- **Formal Definition:** We employ computational techniques, both extractive and abstractive, to create summaries. This dual approach ensures both precision and coherence in summarizing complex academic papers.

Motivation

- **Accessibility:** Making complex scientific texts easily understandable
- **Efficiency in Research:** Saving time for scholars and students
- **Knowledge Synthesis:** Condensing vast information for better knowledge graph management
- **Cutting-Edge Tech:** Applying the latest NLP innovations hands-on

Connections with Course Concepts

- Text Classification and Sentiment Analysis
- Fundamental NLP Models
- Advanced Summarization Algorithms
- Sequence-to-Sequence and Attention Models
- Transformers and T5 Model
- Rouge Score Performance Evaluation

Published Approaches

- **SummaRuNNer:** An RNN-based model for extractive summarization focusing on key sentence selection, emphasizing transparency in summarization
- **Mathematical Modeling:** A paper highlighting the contrast between abstractive and extractive summarization, offering a structured mathematical approach with indicator vectors
- **Attentional Encoder-Decoder RNNs:** Explores abstractive text summarization, addressing challenges in keyword modeling and sentence-to-word hierarchy

Data

- **Data Source:** Utilized a dataset consisting of the top 1000 most cited papers from the ACL Anthology Network (AAN), focusing specifically on the papers with annotated summaries
- **Evaluation Metric - ROUGE**: Measures the quality of summaries by comparing them to reference summaries
    - ROUGE (Recall-Oriented Understudy for Gisting Evaluation) assesses overlap in n-grams, word sequences, and word pairs between the generated and reference summaries.
    - Focuses on coherence, informativeness, and readability of the generated summaries.

Baseline Model Performance

- **Simple baseline:** extracts the first few sentences, assuming these sentences capture the essence or key points of the paper
    - ROUGE-L score achieved was approximately 15%
- **Strong baseline:** using the T5 (Text-to-Text Transfer Transformer) model to transform text summarization into a text-to-text problem
    - ROUGE-L score achieved was approximately 22%

Enhancing Text Summarization with T5 Model

- Advanced preprocessing strategy targeting the starting segments of the texts
- Fine-tuning within a PyTorch Lightning framework
- Training parameters
  - Model Name: ’t5-base’
  - Text Length: 1024 tokens
  - Summary Length: 512 tokens
  - Batch Size: 5
  - Epochs: 10
- ROUGE-L score achieved is 62.85%