Generative AI for Video-to-Text Summarization

Ethan Villalovoz, Fernando Medina, Luke Flock, Tomer Zangi

Background and Motivation

Background: Online video content is growing exponentially, making it difficult for users to find content that has the information they need.

Motivation: To develop solutions that make online videos more searchable and accessible through video summaries.

Why Now? Recent advancements in open-source AI models (e.g., FLAN-T5, BART) make this very possible.

Problem Statement

Instructional videos often lack concise and structured summaries that users can quickly reference.

While generative AI has proven effective in text-based summarization, applying these methods to video data is non-trivial due to:

- The unstructured nature of video captions.
- The need to align captions with task descriptions accurately.
- Computational constraints, particularly for multimodal integration.

Impact: Users are left with the time-consuming task of manually extracting in

formation from long video content.

Objectives

- Align video captions with task descriptions from the HowTo100M dataset.
- Fine-tune advanced AI models like FLAN-T5 and BART.
- Evaluate performance using standard metrics like ROUGE and BLEU.

Previous Objective (scrapped due to limitations with computation, time, etc)

- Multimodal learning with visual and audio features combined

Dataset Overview

Dataset: HowTo100M (over 1.2 million video-caption pairs).

Note: We only used around 25,000 due to computation and time constraints

Key Features:

- Video metadata.
- Task descriptions from WikiHow.
- Time-aligned captions.

Methodology

Preprocessing Pipeline:

- Mapped task IDs to descriptions.
- Parsed captions and removed noise.
- Created a structured dataset for training.

Fine-Tuning:

- Adapted FLAN-T5 and BART for summarization.
- Used captions as input and task descriptions as output.
- Split dataset into training (80%), validation (10%), testing (10%)

HowTo100M Preprocessing pipeline

- Align Metadata: Merged task descriptions with video IDs.

	video_id	category_1	category_2	rank	task_id	task_id	task_description
0	nVbIUDjzWY4	Cars & Other Vehicles	Motorcycles	27	52907	52907	Paint a Motorcycle
1	rwmt7Cbuvfs	Cars & Other Vehicles	Motorcycles	99	52907	52907	Paint a Motorcycle
2	HnTLh99gcxY	Cars & Other Vehicles	Motorcycles	35	52907	52907	Paint a Motorcycle
3	RAidUDTPZ-k	Cars & Other Vehicles	Motorcycles	10	52907	52907	Paint a Motorcycle

- Parse Captions: Structured time-aligned text from caption.json.

	video_id	task_description	captions
0	nVbIUDjzWY4	Paint a Motorcycle	{'start': [13.64, 15.86, 20.6, 23.96, 26.36, 2...
1	rwmt7Cbuvfs	Paint a Motorcycle	{'start': [1.8, 6.32, 7.32, 10.86, 13.28, 15.6...

- Clean Data: Removed stop words, redundancies.
- Output: Structured CSV for downstream tasks.

Overview of Fine-Tuning

Objective:

To train a generative AI model (FLAN-T5) on the HowTo100M dataset to produce concise video summaries.

Key Steps:

- Load and preprocess data.
- Prepare tokenized datasets.
- Define and implement a training loop.
- Monitor performance via validation.
- Evaluate with ROUGE and BLEU metrics.

Fine-Tuning Dataset Preparation

Dataset: Using the preprocessed HowTo100M dataset.

Input (Source Text): Video captions concatenated into plain text (e.g., "Primed motorcycle fenders, applied base coat...").

Output (Target Summary): Task descriptions from WikiHow (e.g., "Paint a Motorcycle").

Statistics:

- Training set: 80% (19,823 samples).
- Validation set: 10% (2,476 samples).
- Testing set: 10% (2,478 samples).

Model and Training Setup

Models Used: FLAN-T5 and BART

Training Details:

- Multi-GPU setup with 4 NVIDIA RTX A6000 GPUs (resources shared with others so sometimes there were computation problems).
- Optimizer: AdamW with a learning rate of 5e-5.
- Loss Function: Cross-entropy loss.
- Epochs: 3, Batch Size: 16.

Training Process

Training Phase:

- Model learns input-output relationships using training dataset.
- Forward pass computes loss, backpropagation updates parameters.

Validation Phase:

- Evaluates model performance on unseen data after each epoch.
- Tracks validation loss to prevent overfitting.

Checkpointing:

- Saves model after each epoch for recovery and tuning.

Evaluation Methods (BLEU)

Definition: Measures how well the generated text matches reference text using n-gram precision.

BLEU focuses more on the similarity of the generated and reference text instead of the content recovery and relevance that ROUGE looks for.

Components:

- 1-gram: Individual word matches.
- 2-gram: Two-word phrase matches.
- 3-gram and 4-gram: Longer sequences (fluency and coherence).

Evaluation Methods (ROUGE)

We are evaluating precision, recall, and F-measure

Definition: Measures the overlap between generated text and reference text.

Recall is the portion of the reference summary that was recovered by the AI.
Precision is the portion of the AI generated summary that is relevant.

Components:

- ROUGE-1: Unigram overlap (word-level match).
- ROUGE-2: Bigram overlap (two-word sequences).
- ROUGE-L: Longest common subsequence (sentence structure match).

Flan-T5 Sample Summary Outputs

Example 1:

Generated Summary: "Make a Grass Root Beer."

Reference Summary: "Paint a Motorcycle."

Example 2:

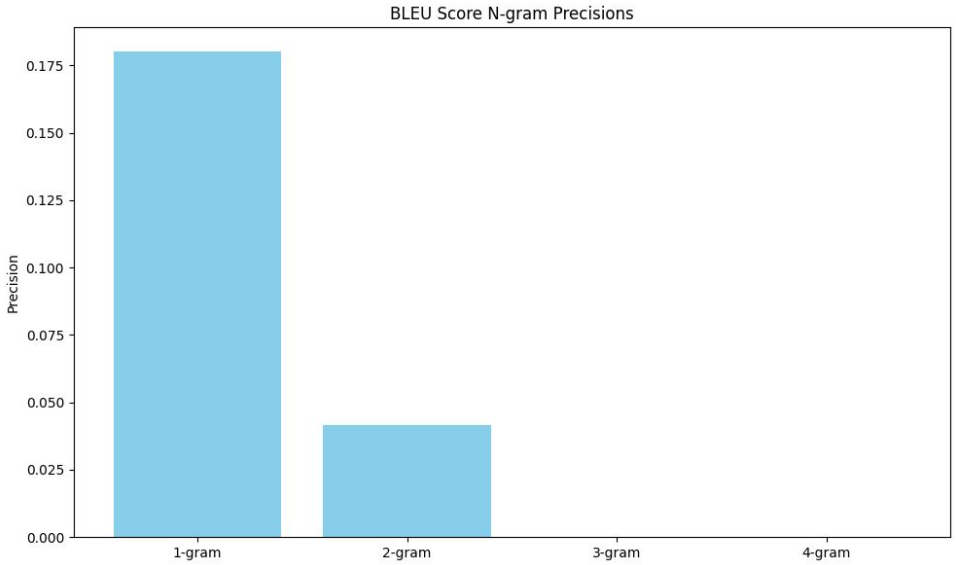
Generated Summary: "Assemble a Bicycle Frame."

Reference Summary: "Install Motorcycle Handlebars."

Insights:

Summaries aren’t the best, either irrelevant or too generic

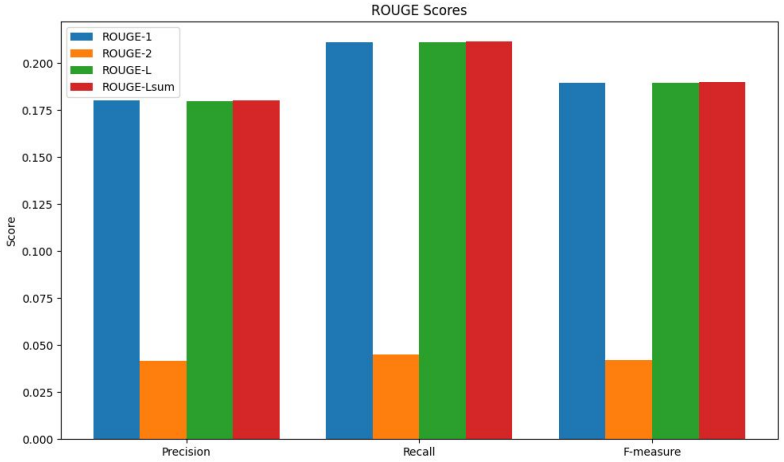
Flan-T5 BLEU Evaluation Results



Key Observations:

- BLEU measures precision across n-grams (1-gram to 4-gram).
- Highest precision observed for 1-gram (18%), indicating the model captures individual words accurately.
- Sharp decline for 2-gram and higher, suggesting challenges in generating coherent multi-word sequences.

Flan-T5 ROUGE Evaluation Results



Graph Insights:

ROUGE-1 and ROUGE-L achieved higher scores (~0.18–0.21).

ROUGE-2 (bigrams) was significantly lower (~0.04), indicating challenges in capturing multi-word patterns.

BART Sample Summary Outputs

Example 1:

Generated: "Make a Duct Tape Bracelet."

Reference: "Make Crawfish Bisque."

Example 2:

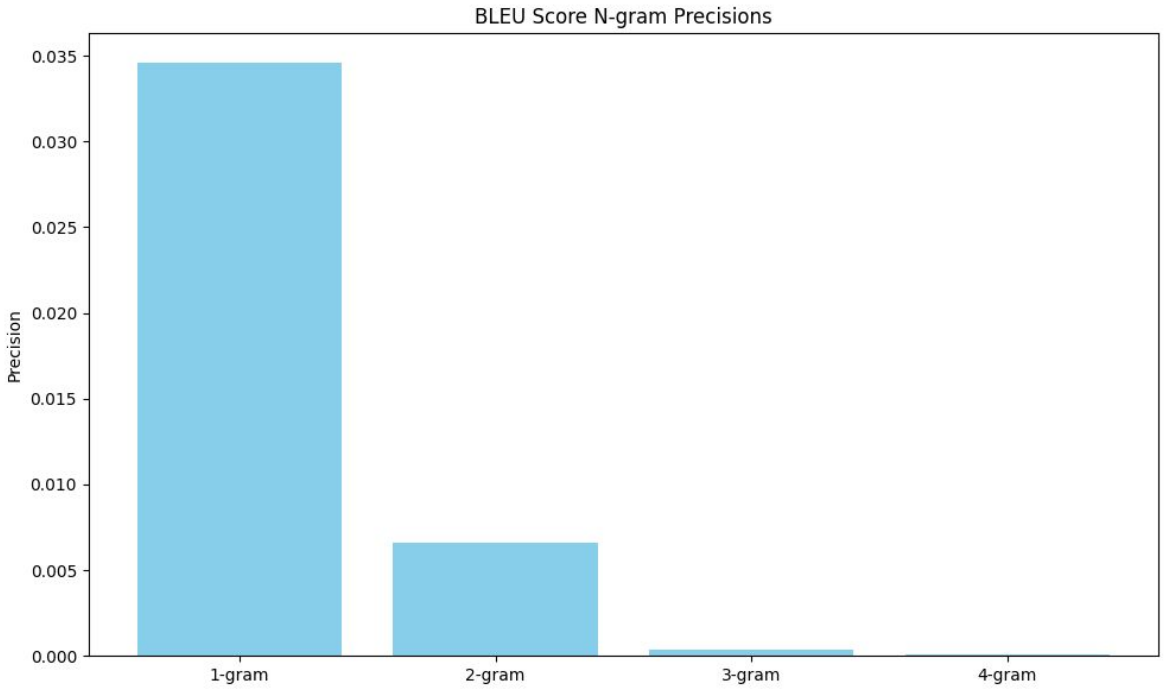
Generated: "Make a Plastic Bottle Bracelet."

Reference: "Measure Vital Signs During First Aid."

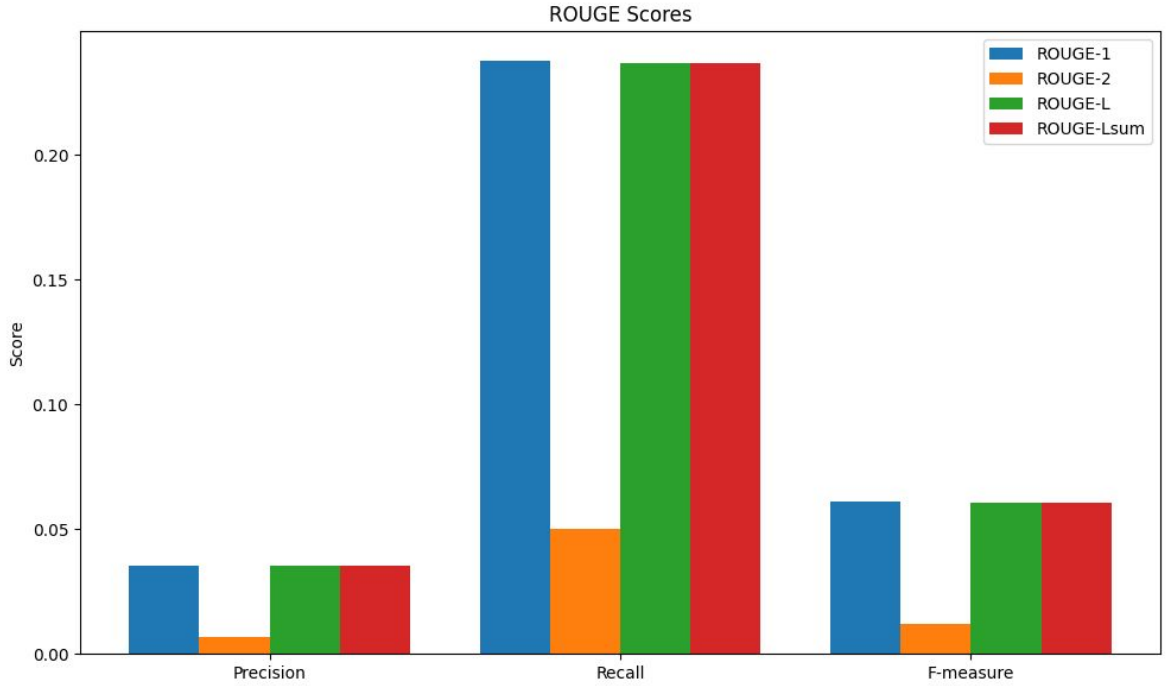
Insights:

Irrelevant outputs despite training efforts.

BART BLEU Evaluation Results



BART ROUGE Evaluation Results



Comparison

BART Strengths:

- Effective token-level precision (high ROUGE-1, BLEU 1-gram).
- Captures simple sentence patterns reasonably well.

BART Weaknesses:

- Lower overall ROUGE scores reflect challenges in contextual understanding.
- Struggles with tasks requiring high-level comprehension (e.g., summary coherence).

FLAN-T5 vs. BART:

- FLAN-T5 excels in recall and coherence, making it more suitable for tasks that require summarizing lengthy, complex content while maintaining global structure.
- BART performs better at precise token and n-gram matching, making it advantageous for tasks that prioritize lexical accuracy and concise summaries.

Conclusion

Moderate Success in Summarization:

- Both FLAN-T5 and BART demonstrated the ability to generate summaries, but the outputs were often far from ideal.
- BLEU and ROUGE scores highlight these limitations, with low n-gram precision and recall reflecting poor lexical and contextual accuracy.

Thanks everyone for watching