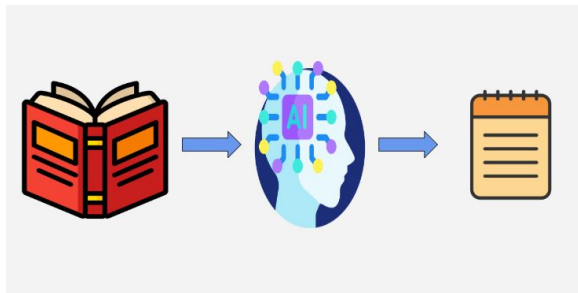


Summarization for News Data

Ramy Qranfal and Jake Carter

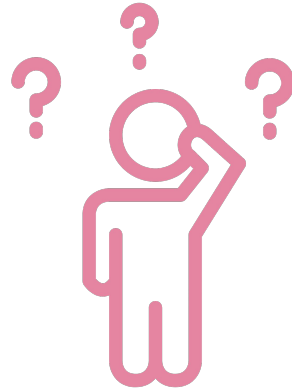
Introduction

- AI-generated summaries often lack clarity and a natural, human-like writing style
- Our solution is to fine-tune a model according to human-written summaries, hopefully resulting in a more human-like output
- The dataset used includes CNN and Dailymail articles with author-written summaries
- Our project goal is to provide concise and readable summaries of the news for people interested in the news but have little time to keep up with it



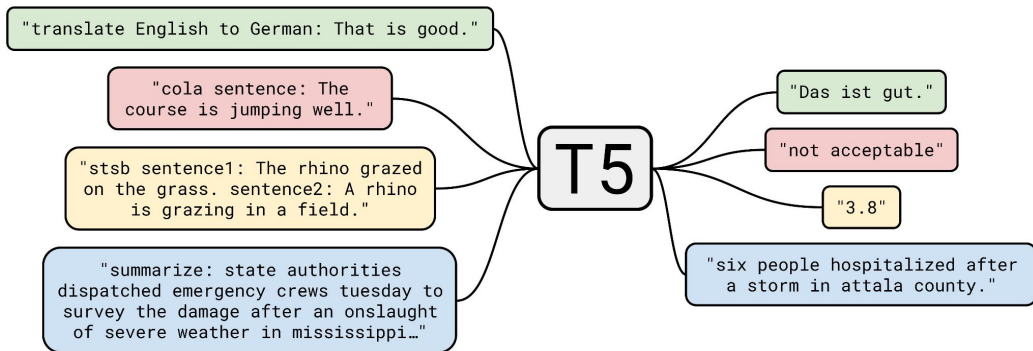
Problem

- People are often interested in keeping up with current events, but this can be daunting
 - Many different news outlets with time-consuming shows and articles that the working person might struggle to keep up with
- AI-generated summaries often lack clarity and a natural, human-like writing style
 - This lack of a human personality could put many people off of AI news as it appears stale and lifeless



Pretrained LLM

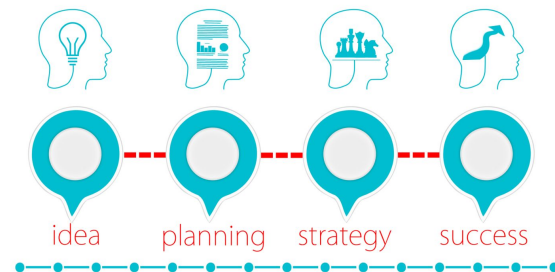
- Our chosen model is Google's T5-base
- The T5 models are designed for text-to-text tasks. This makes them ideal for summarizations
- The unsupervised pretraining of the T5 reduces overfitting and enhances task adaptability
- The T5-base model was chosen for its balance between performance and resource cost
 - T5-small is less resource-intensive but results in lower performance
 - T5-large has higher performance at the cost of more computational resources



Methodology

Steps while utilizing the pretrained T5-base model as the base:

1. Loaded dataset and T5's tokenizer
2. Sampled the large dataset for fine tuning
3. Tokenize text
4. Create a scoring system using the ROGUE score
5. Set parameters and training weights
6. Send the model, scoring metrics, and dataset to the trainer
7. Begin training to fine-tune the model



Results (Final ROUGE Scores)

- ROUGE-1: 26.38% (Originally 23.47%)
 - Shows the model grasps key vocabulary and context
- ROUGE-2: 12.28% (Originally 10.77%)
 - Demonstrates the model's ability to be logical
- ROUGE-L: 18.35% (Originally 16.33%)
 - Shows the model's ability to retain and structure sentences
- ROUGE-Lsum: 18.38% (Originally 16.36%)
 - Measurement for how the model keeps the overall structure of the content

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	RougeL	RougeLsum	Gen Len
1	0.770800	0.830198	0.234700	0.107700	0.163300	0.163600	53.630000
2	0.723600	0.823444	0.255100	0.118400	0.177100	0.177500	57.932000
3	0.727000	0.822413	0.261700	0.120800	0.181900	0.182400	60.128000
4	0.730100	0.821479	0.261400	0.120800	0.182100	0.182300	59.024000
5	0.730800	0.821524	0.263800	0.122800	0.183500	0.183800	59.633000
6	0.730400	0.821548	0.263900	0.122100	0.183500	0.183700	59.929000

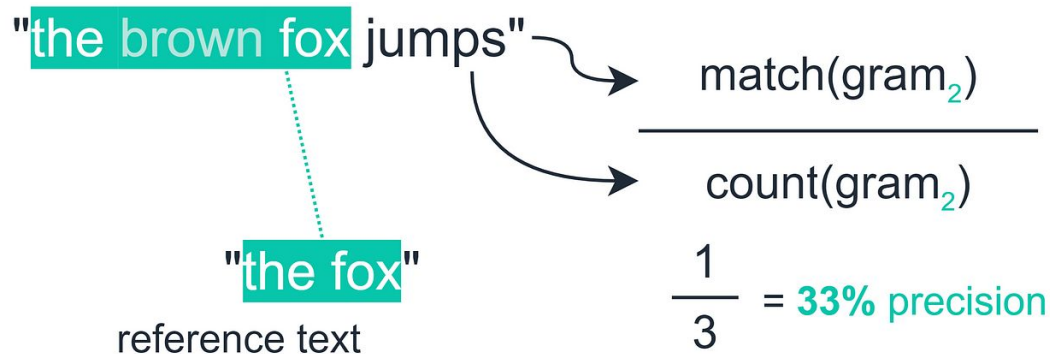
Accuracy

- ROUGE scores steadily increased across epochs
 - Improved unigram and bigram coverage and with stronger summary structure
- Generated summaries became longer and more detailed as training progressed (average length went from 53 to 60 tokens).
- Strong signs of improvement based on 10,000 samples on the T5-base model

"the brown fox jumps"

reference text

"the fox"

$$\frac{\text{match}(\text{gram}_2)}{\text{count}(\text{gram}_2)} = \frac{1}{3} = 33\% \text{ precision}$$


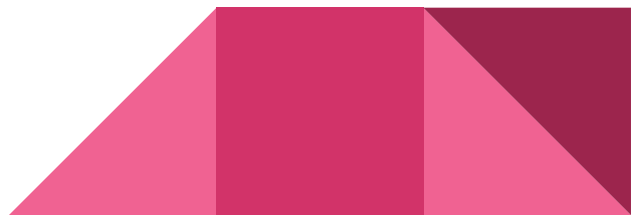
Examples

GENERATED SUMMARY:

Kuku Kube was created by Canada-based Network365 and is available for free on Facebook, Android, iOS and on desktop browsers. Players get a point for every correct square identified, but if they click or tap the wrong square they lose a point. Scores lower than 11 are poor, scores between 15 and 20 is 'lower than average', 21 to 30 is considered normal or average, and a score higher than 31 means your eyesight is considered great.

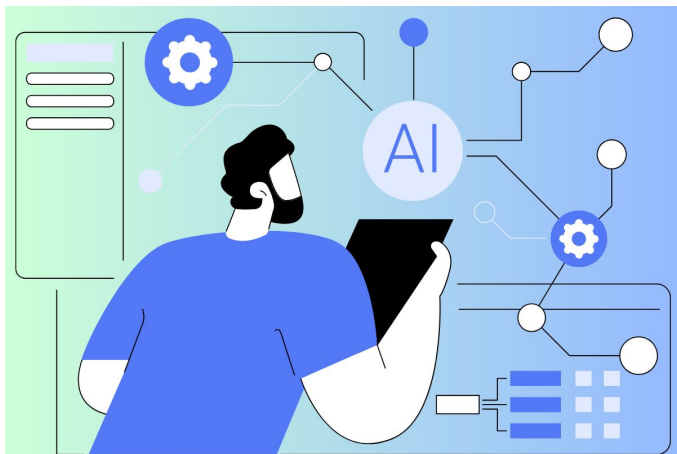
REFERENCE SUMMARY:

Free app is available on Facebook, Android, iOS and on desktop browsers. It starts with four squares and asks you to identify the different shade. Board grows to up to 81 squares and differentiation is subtle each time. And a score of 31 or above is considered a sign of 'great eyesight'.



Analysis

- T5-base shows a strong ability to produce human-like output on a limited dataset
- ROUGE scores show a continually improving similarity to human output
- Results could be improved on the T5-large model with a larger size of training samples
- The model showed solid improvement in the first 3 epochs, while the final 3 epochs had minimal gains indicating convergence.



Conclusion

- Fine-tuned the T5-base model using human-written summaries of articles
 - Accomplished via smart sampling, parameter adjustments, and preprocessing.
- Steadily improving ROUGE scores show this model can achieve strongly human-like summaries



Thank you!

