

Summarizing the News

Deep Generative Models

Mowafak Allaham, Julia Barnett

Computational Journalism



Computational Journalism

- Newsrooms around the world have been using more and more *computational methods*:
 - Information gathering
 - Production
 - Sensemaking
 - Distribution
 - Audience consumption behavior
 - Computational news discovery



Newsroom Summarization

- Over 5,000 articles are published online per day (Hamborg, Meuschke, and Gipp 2018)
- 50% of US adults get news media from social media (Leidke & Matsa, 2022)
- To keep abreast of the information out there, solution: **news summarization tools**
 - Useful for journalists
 - Daily briefing of news
 - Quickly know whether an article is relevant

Our formal research objective:

1. Develop a ***proof-of-concept tool*** using a deep generative model that can help journalists ***summarize news articles*** from a range of news sources and across news categories
2. Highlight ***potential risks*** that may occur from these summarizations—especially in regards to conspiracy claims in low credible news sources

Text Summarization Tools and Methods:



Models



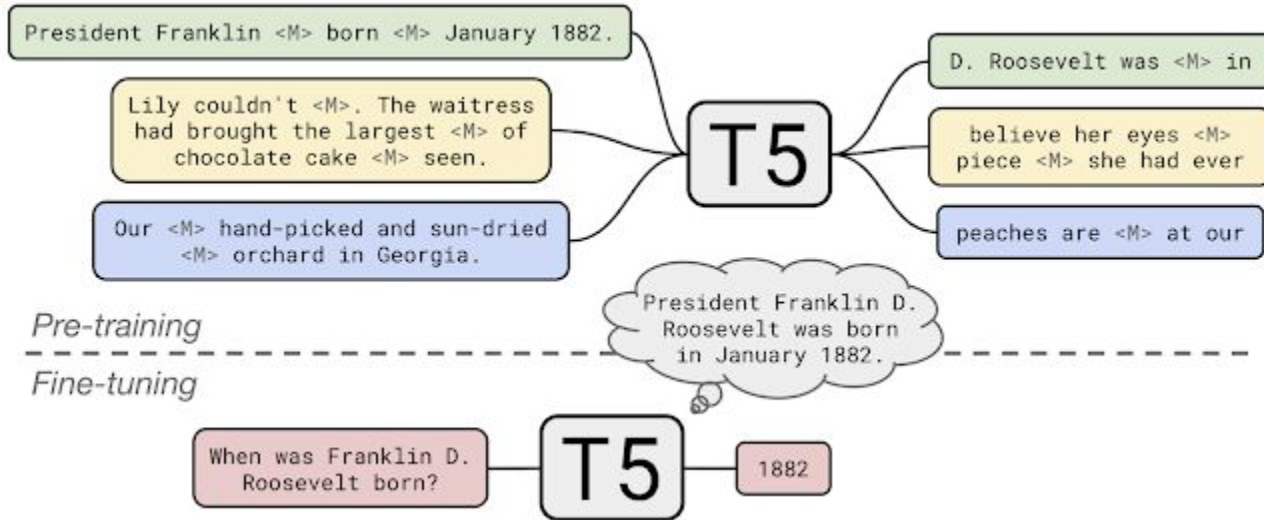
- **GPT3 (OpenAI)**
 - *Drawbacks:* enormous, expensive to use
- **Mini versions of GPT3:**
 - **GPT-Neo**
 - **GPT-J**
 - *Drawbacks:* couldn't get these models to work for text summarization. However, they do work for a general text generation task based on prompt.
- **PEGASUS:**
 - *Drawbacks:* Did not yield interpretable summaries
- **T-5:**
 - Our choice!
 - Most affordable and versatile model for text



Text-To-Text Transfer Transformer (T5 Model)

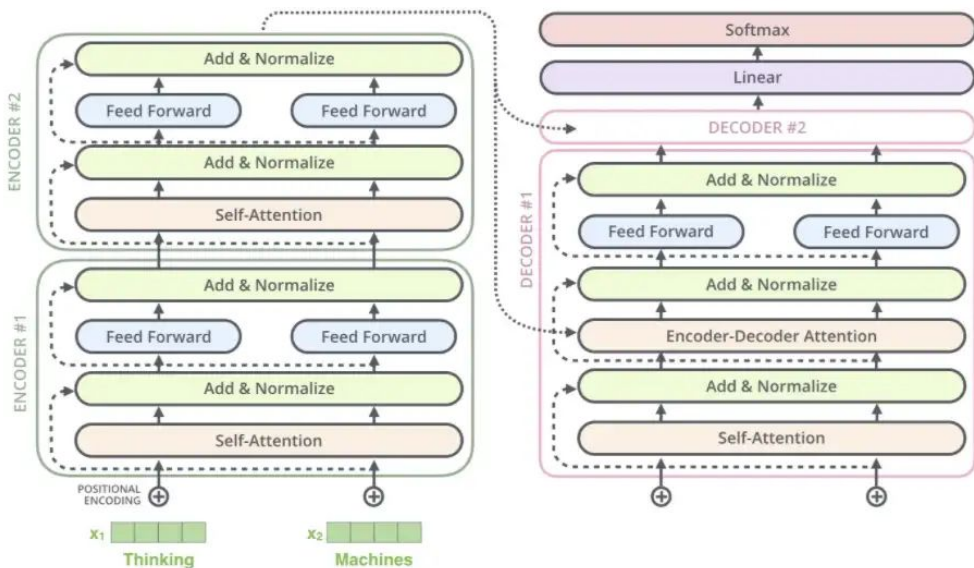
Motivation behind T5 Model

- Provide a unified **text-to-text framework** (input and output are text)
- Combine successful **transfer learning techniques** in NLP models into a single model
- Train a language model on **“clean” data**: Colossal Clean Crawled Corpus (C4)



T5 Model (architecture)

- **Transformer-based model** which includes an encoder-decoder transformer
- The encoder-decoder transformer is inspired by transformers with **self attention architecture** (Vaswani et al., 2014)
- Placed layer normalization outside residual path
- Removed **Layer Norm bias**
- Used different positional embedding
 - Variation of relative positional embedding
 - (Still not clear to us!)
- Offers the following model variants:
 - Base (220 million parameters)
 - Large (770 million parameters)
 - 3B (3 Billion parameters)
 - 11B (11 Billion parameters)



T5 Model - Training Data

- Models are trained on lines that **ended in a terminal punctuation mark** (. ? ! .")
- Discarded:
 - Pages with < 5 sentences
 - Sentences with < 3 words
 - Pages containing any word on the “List of Dirty, Naughty, Obscene or Otherwise Bad Words” (Shutterstock 2020)
- Deduplicated the data set by discarding all but one of any three-sentence span occurring more than once in the data set

T5 Model downstream tasks – Translation



T5 Model downstream tasks – Trivia

question 1

Who wrote the epic 1869 novel War and Peace

You

SUBMIT

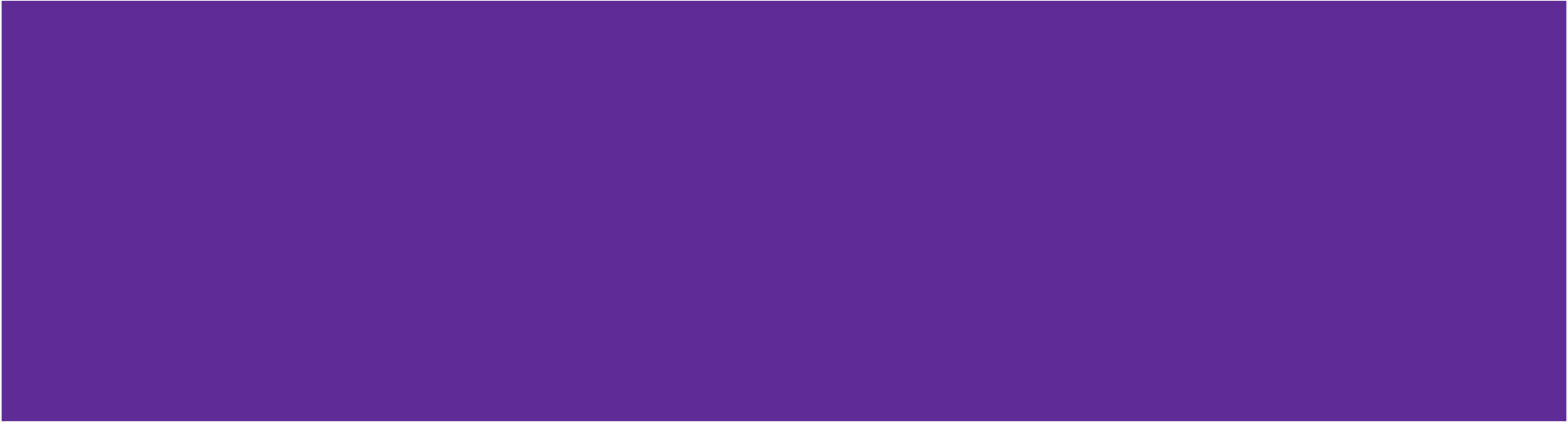
0

T5

HMM...

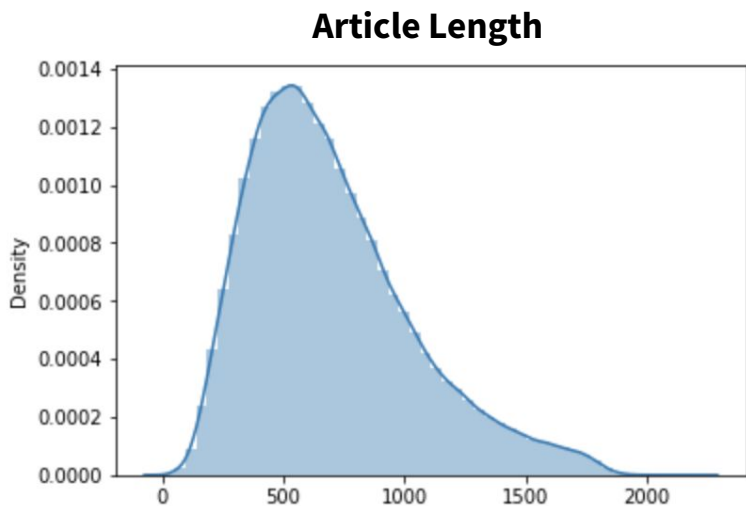
0

Tuning the Model + Data

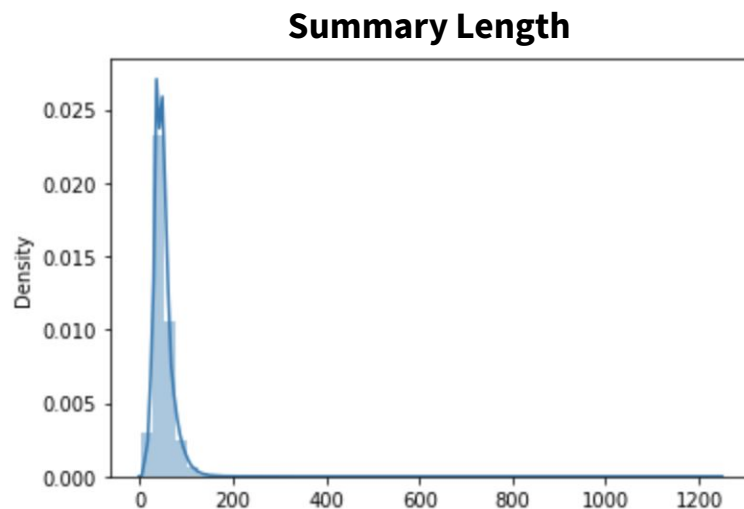


Fine Tuning the Model: CNN and the Daily Mail

- Article + summary pairs from CNN and the Daily Mail (See, Liu, and Manning 2017)
 - Training set: 229,690
 - Test set: 57,423



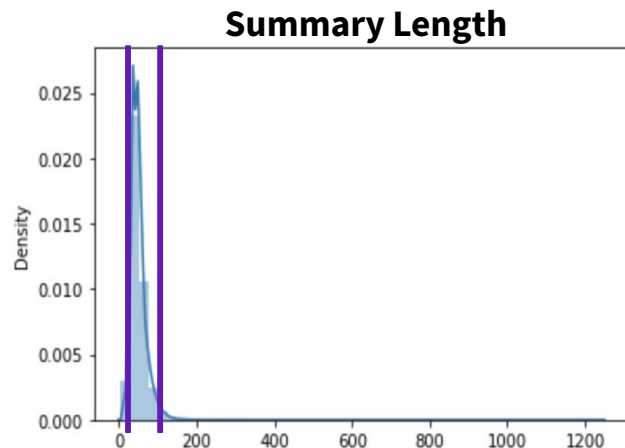
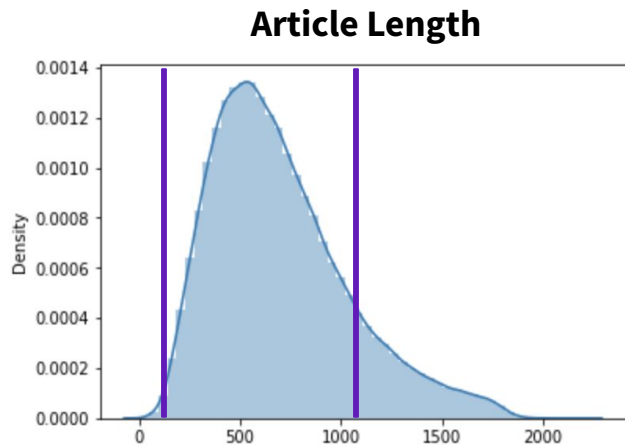
$\mu = 698$; $\sigma = 338$; median = 638.0



$\mu = 49.5$; $\sigma = 20.8$; median = 46.0

Training decision

- For fine-tuning the model:
 - Prefixed all articles with “**summarize:**” prompt
 - Max article length = 1024 (*truncating enabled*)
 - Max summary length = 100 (*truncation enabled*)
- Summarization generation:
 - Min summary length = 50
 - Max summary length = 100
- Trained the model on 5 steps with
 - least cross entropy loss achieved: 0.07478



Summarizing low credible sources

- 185,080 articles and blog posts (between 2017 and 2022) substantially discussing climate change
- Identified presence of false or misleading CC claims at the paragraph level of analysis (using the CARDS (Computer-assisted Recognition of CC Denial and Skepticism))
- Generated summaries for **150 low credible articles** (25 per year 2017-2022):
 - Summary length = 50-100 words
 - Max length to encode is 1024 characters (with truncation enabled)
 - Max sequence length is 1024
 - The model caps on 512 and a batch size of 128 ~ 65K tokens in a single step
 - Each token is assigned a unique ID to the input sequence

17 Claim Types

- Ice/permafrost/snow cover isn't melting
- We're heading into an ice age/global cooling
- Weather is cold/snowing
- **Climate hasn't warmed/changed over the last (few) decade(s)**
- Sea level rise is exaggerated/not accelerating
- Extreme weather isn't increasing/has happened before/isn't linked to climate change
- It's natural cycles/variation
- **There's no evidence** for greenhouse effect/carbon dioxide driving climate change
- Climate sensitivity is low/negative feedbacks reduce warming
- Species/plants/reefs **aren't showing climate impacts**/are benefiting from climate change
- CO2 is beneficial/not a pollutant
- Climate policies (mitigation or adaptation) are harmful
- **Climate policies are ineffective/flawed**
- Clean energy technology/biofuels won't work
- People need energy (e.g. from fossil fuels/nuclear)
- **Climate-related science is unreliable**/uncertain/unsound (data, methods & models)
- Climate movement is unreliable/alarmist/corrupt

Results



Qualitative Analysis

- ❖ Main types of summaries
 - Accurate summarization:
 - Perpetuated the claim (22%)
 - Removed the claim entirely (78%)
 - Poor quality/not intelligible

Qualitative Analysis

❖ Main types of summaries

- Accurate summarization:
 - **Perpetuated the claim**
 - Removed the claim entirely
- Poor quality/not intelligible

“ global elites are pivoting on climate change. there will be other hysterics, but no real political push on the climate front ever again – just as war on terror did in 2004!

”

Original article: arguing both COVID and climate change were political hoaxes.

Qualitative Analysis

❖ Main types of summaries

- Accurate summarization:
 - Perpetuated the claim
 - **Removed the claim entirely**
- Poor quality/not intelligible

“ environmentalists want epa inspector general to investigate whether pruit violated internal policies. sierra club attorney: pruits comments contradict basic scientific findings of research on climate change, not contrary ones from studies - 'obviously false' former vice president al gore said comments were "independent and unfounded" ”

Original article: endorsed Pruitt's false claims of climate change.

Qualitative Analysis

❖ Main types of summaries

- Accurate summarization:
 - Perpetuated the claim
 - **Removed the claim entirely**
- Poor quality/not intelligible

“ last year's atlantic hurricane season produced 10 storm surge events, including six major hurricane types. the 2018 forecast has three or four storm potential impacts on the u-pacific coast if they do hit the u us ; but that'll be less than last summer' average for this season compared to last august eve a few storm seasons ago dubbed "the worst of our history ”

Original article: claimed that hurricanes were slightly worse than normal, but not a result of climate change and would revert back to neutral soon.

Qualitative Analysis

❖ Main types of summaries

- Accurate summarization:
 - Perpetuated the claim
 - Removed the claim entirely
- **Poor quality/not intelligible**

“ a new study shows that plants are to blame for warming, not man's. the theory is that man'd produce more co2 to feed people and poison them - but it doesnt work this way if we increase carbon dioxide levels in our cars or on highway roads... it'll be madness. ”

You can tell the claim is there, but the summary doesn't quite make sense.

Future Work

- Analyze other types of content perpetuated or removed by summaries
 - Political bias
 - Other dangerous conspiracy theories
- Examine how the credibility of the news organization affects the summaries
- Train a model to classify whether the claims are present or not
- Extend to different data sources
- Add more user specificity to the summarization tool (“I want a summary about this part of the article”)