nGPT, BART and PEGASUS: A Comparative Study

Presented by: Shishir Kallapur, Sanjiv Shrirang Joshi and Pratheesh

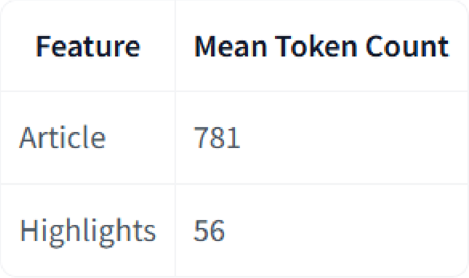
For: CS6120 Semester Project Fall 24

Objectives

- Assess nGPT's claims of achieving 4-20x faster training times and improved stability for Large Language Models.
- Compare nGPT against BART and PEGASUS on several aspects.
- Analyze each model's efficiency through various inference metrics.
- Evaluate summary generation using ROUGE scores and human annotation.

CNN / DailyMail Dataset

- News articles
- CNN from April 2007 to April 2015
- DailyMail from June 2010 to April 2015
- **Reason for use**: Popular benchmark for text summarization
- The sample splits:
  - 287k training
  - 13k validation
  - 11k test
- **Subset**: 15k samples to train, 1.5k samples to validate.



Feature	Mean Token Count
Article	781
Highlights	56

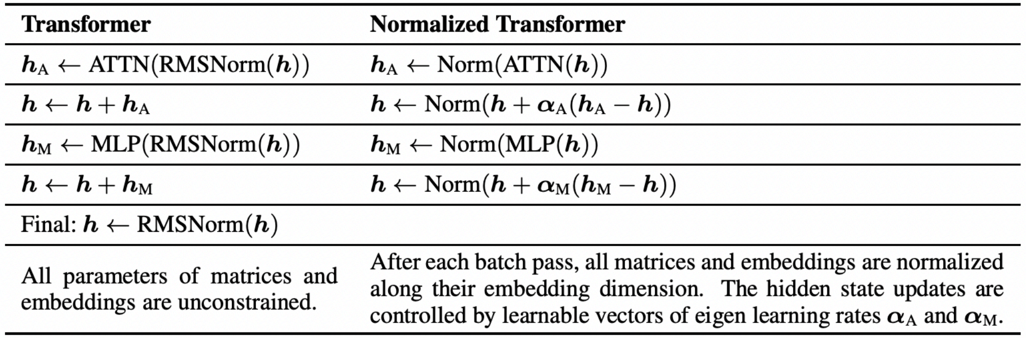
{'**id**': '0054d6d30dbcad772e20b22771153a2a9cbeaf62',
'**article**': '(CNN) -- An American woman died aboard a cruise ship that docked at Rio de Janeiro on Tuesday, the same ship on which 86 passengers previously fell ill, according to the state-run Brazilian news agency, Agencia Brasil……
''**highlights**': 'The elderly woman suffered from diabetes and hypertension, ship's doctors say .\nPreviously, 86 passengers had fallen ill on the ship, Agencia Brasil says .'}

Data Schema

CNN Dailymail Dataset: https://huggingface.co/datasets/abisee/cnn_dailymail

nGPT: Normalized Transformer with Representation Learning on the Hypersphere

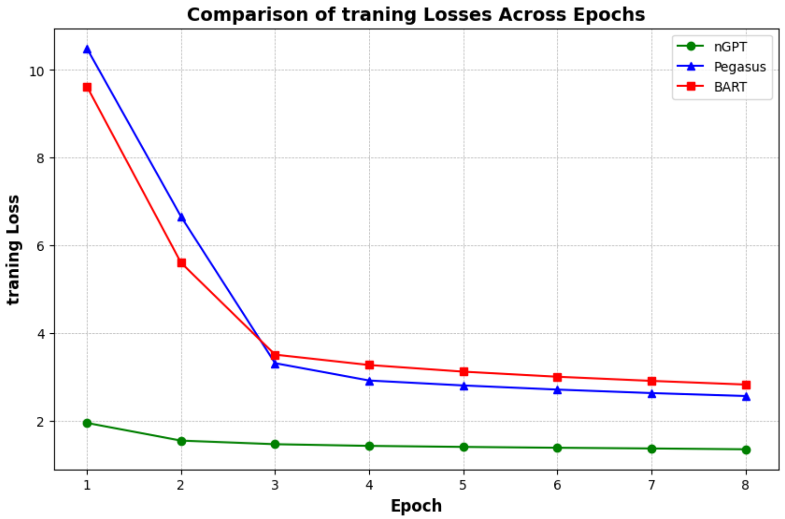
- **Key Features**:
  - Removes LayerNorm/RMSNorm; normalizes matrices after training steps.
  - Adjusts softmax scaling factor and embedding rescaling.
  - No weight decay; restructured intermediate states and logits scaling.
  - Optimized hyperparameters for stability and convergence.
- **Comparison Models**: BART and PEGASUS

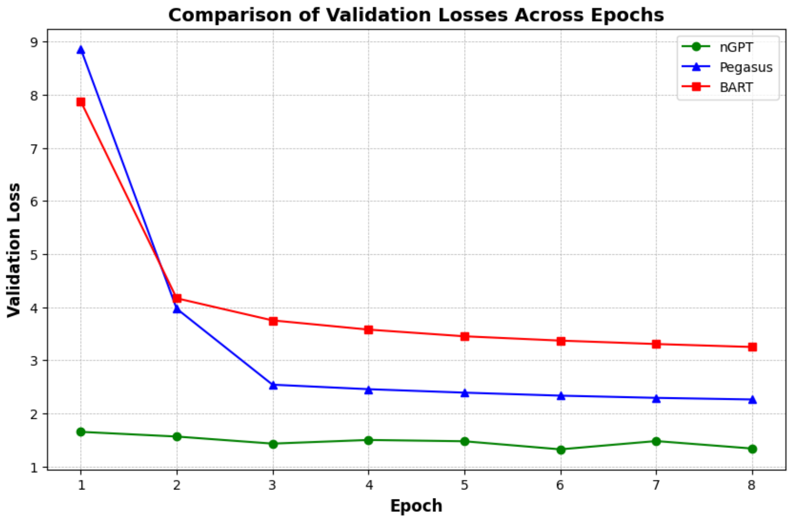


Transformer	**Normalized Transformer**
$\boldsymbol{h}_{\mathrm{A}} \leftarrow \mathrm{ATTN}(\mathrm{RMSNorm}(\boldsymbol{h}))$	$\boldsymbol{h}_{\mathrm{A}} \leftarrow \mathrm{Norm}(\mathrm{ATTN}(\boldsymbol{h}))$
$\boldsymbol{h} \leftarrow \boldsymbol{h} + \boldsymbol{h}_{\mathrm{A}}$	$\boldsymbol{h} \leftarrow \mathrm{Norm}(\boldsymbol{h} + \boldsymbol{\alpha}_{\mathrm{A}}(\boldsymbol{h}_{\mathrm{A}} - \boldsymbol{h}))$
$\boldsymbol{h}_{\mathrm{M}} \leftarrow \mathrm{MLP}(\mathrm{RMSNorm}(\boldsymbol{h}))$	$\boldsymbol{h}_{\mathrm{M}} \leftarrow \mathrm{Norm}(\mathrm{MLP}(\boldsymbol{h}))$
$\boldsymbol{h} \leftarrow \boldsymbol{h} + \boldsymbol{h}_{\mathrm{M}}$	$\boldsymbol{h} \leftarrow \mathrm{Norm}(\boldsymbol{h} + \boldsymbol{\alpha}_{\mathrm{M}}(\boldsymbol{h}_{\mathrm{M}} - \boldsymbol{h}))$
Final: $\boldsymbol{h} \leftarrow \mathrm{RMSNorm}(\boldsymbol{h})$	
All parameters of matrices and embeddings are unconstrained.	After each batch pass, all matrices and embeddings are normalized along their embedding dimension. The hidden state updates are controlled by learnable vectors of eigen learning rates $\boldsymbol{\alpha}_{\mathrm{A}}$ and $\boldsymbol{\alpha}_{\mathrm{M}}$.

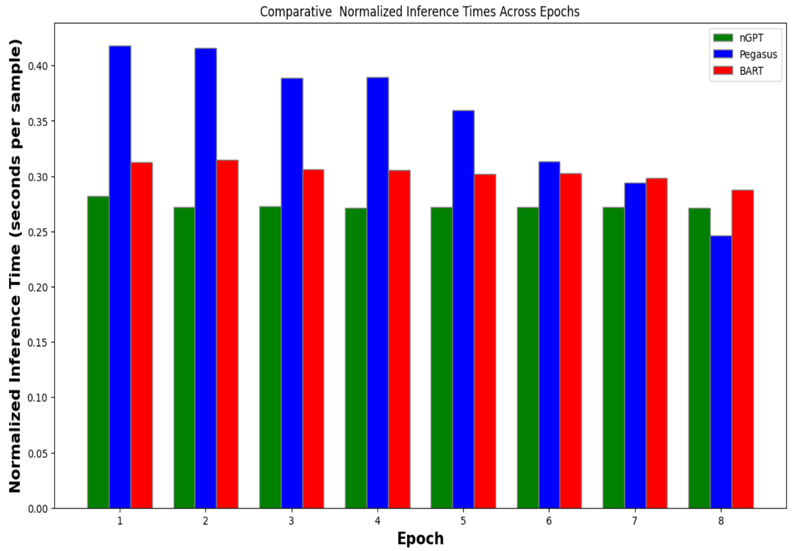
nGPT: https://arxiv.org/abs/2410.01131

Experimental Results


Comparison of traning Losses Across Epochs


Comparison of Validation Losses Across Epochs

Note: the changes in the values for training and validation losses for nGPT is dampened due to scaling of other graphs

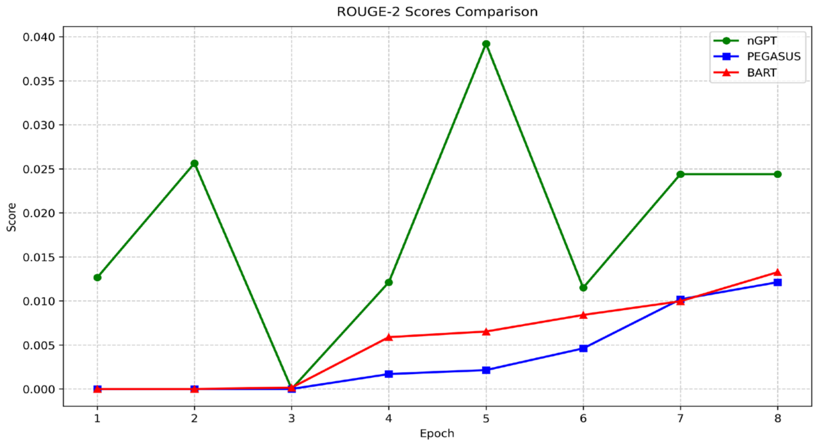

Comparative Normalized Inference Times Across Epochs

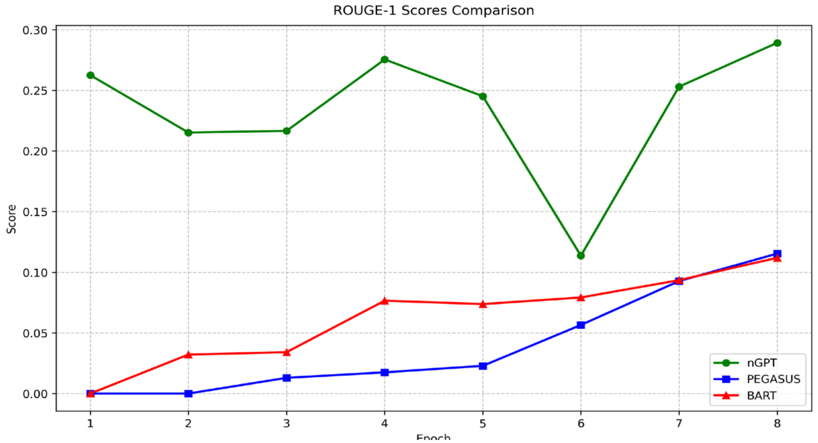
- Key takeaways
  - Lower training loss, capturing the patterns of the data
  - Lower inference times, sampling summaries better and efficiency
  - Early consistent training loss, shows better stability
  - Early decrease in validation, shows better generalization on unseen data

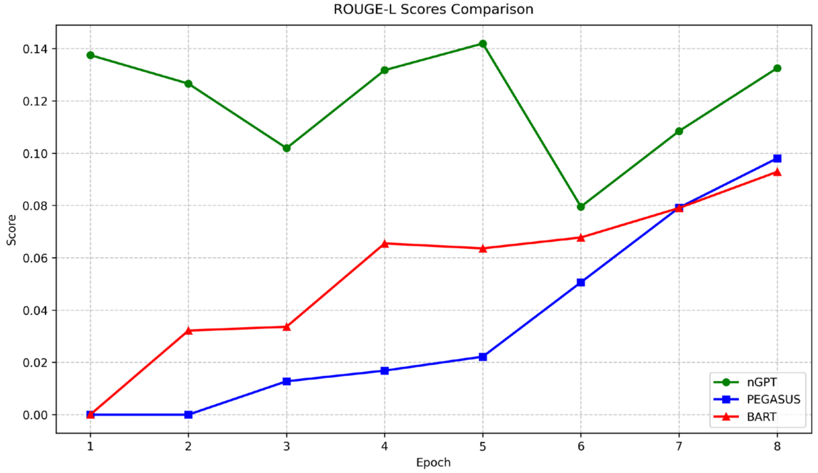
ROUGE Scores

- nGPT outperforms PEGASUS and BART, achieving ROUGE-1 scores up to 0.29 (29%), ROUGE-2 scores up to 0.039 (3.9%), and ROUGE-L scores up to 0.14 (14%) compared to the baseline models' lower performance (ROUGE-1: ~11%, ROUGE-2: ~1.2%, ROUGE-L: ~9%)

- Current results show promising improvements in summarization quality, though with observable variability; scaling to larger batches and more training data would likely stabilize performance while maintaining nGPT's efficiency advantages


ROUGE-2 Scores Comparison


ROUGE-1 Scores Comparison


ROUGE-L Scores Comparison

Generated Summaries

(Real Simple) -- Here are five great ways to enjoy your summer. Lazing in a hammock is one of the best ways to spend a summer evening. Best way to cut jeans into shorts -- What better way to declare the start of summer? The key to cutting off jeans is not to go too short too soon. Slip on the jeans and mark the desired length on one leg with chalk……. [Truncated]

nGPT (100M 15K)	BART (100M 15K)	PEGASUS (100M 15K)
Hours the preceding locations along with measurements and the intracrants thereof can tend to be less decreased in perfection, perfection and perfect	Discovery takes place on the summer's leg leg leg with a leg leg . Says it can help help help get back to get away and get back "We're in your way to make you you you like you you," she says .At least one of the world's best way to be able to get back back .	CNN's new album is a new album of the best in the world . He says you can be able to be a good for you't be able . She says you know you can't know what you know what it't make it
	BART (570M all)	PEGASUS (406M all)
	Lazing in a hammock is one of the best ways to spend a summer evening. The best way to catch fireflies is with womanly wiles. Running on the beach can get you into great shape.	Real Simple offers five great ways to enjoy your summer .<n>Best way to catch fireflies -- How? With womanly wiles .<n>The best way to get in and out of a hammock -- Everyone looks good in a hammock .<n>The best way to tie espadrilles -- Apply this lace-up logic .

BART: https://huggingface.co/facebook/bart-large-cnn
PEGASUS: https://huggingface.co/google/pegasus-cnn_dailymail

Limitations

Project's Perspective

- Lack of resources and GPUs on GCP.
- Lack of readily available model implementation for nGPT. Our implementation was not furthered.
- Open-source implementation to make it fit for our chosen dataset.

Model's perspective

- nGPT's autoregressive nature limits its ability to act as a strong abstractive summarizer compared to PEGASUS/BART.
- May cause repetition of tokens.

nGPT-pytorch: https://github.com/lucidrains/nGPT-pytorch

Thank you

Questions?

Link to the presentation

https://youtu.be/FcNyho4Wir8