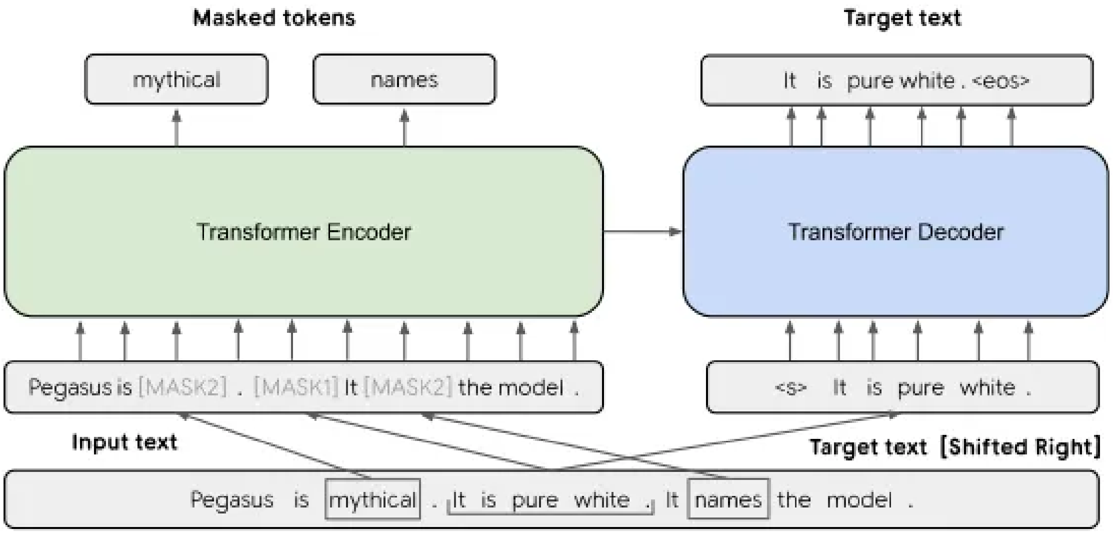
PEGASUS

PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence models) is a transformer-based model developed by Google Research, specifically designed for abstractive text summarization. It employs a self-supervised pre-training objective called Gap Sentences Generation (GSG), where important sentences are removed from a document and the model is trained to generate these sentences from the remaining text. This approach closely mirrors the summarization task, enabling PEGASUS to achieve state-of-the-art performance on various summarization benchmarks.


Masked tokens
mythical
names
Target text
It is pure white . <eos>
Transformer Encoder
Transformer Decoder
Pegasus is [MASK2] . [MASK1] It [MASK2] the model .
<s> It is pure white .
Input text
Target text [Shifted Right]
Pegasus is mythical . It is pure white . It names the model .

1. **Gap Sentences Generation (GSG):** During pre-training, PEGASUS selects and masks entire sentences deemed important (gap sentences) from a

document. The model is then trained to generate these sentences based on the remaining text, effectively learning to summarize.

2. **Sentence Selection Strategies:** To determine which sentences to mask, PEGASUS employs strategies such as random selection, leading sentences, or principal sentences identified through metrics like ROUGE scores. This ensures the model learns to predict content that is truly representative of the document's essence.

3. **Masked Language Modeling (MLM):** In addition to GSG, PEGASUS initially incorporates MLM, where a percentage of tokens in the input are masked and the model learns to predict them. However, studies found that MLM contributed less to summarization performance, leading to its exclusion in the final model.

4. **Pre-training Corpus:** PEGASUS is pre-trained on large corpora such as C4 (Common Crawl) and HugeNews, encompassing a diverse range of topics and writing styles. This extensive training enables the model to generalize well across different domains.

5. **Fine-tuning on Downstream Tasks:** Post pre-training, PEGASUS is fine-tuned on specific summarization datasets like XSum, CNN/DailyMail, and others, allowing it to adapt to the nuances of various summarization tasks.

Applications of PEGASUS

PEGASUS's architecture and training make it highly effective for a range of applications:

- **News Summarization:** Automatically generating concise summaries of news articles, aiding in quick information dissemination.

- **Scientific Paper Summarization:** Condensing lengthy research papers into brief abstracts, facilitating easier comprehension of scientific literature.

- **Legal Document Summarization:** Summarizing complex legal documents to extract key information, useful for legal professionals and laypersons alike.

- **Instructional Content Summarization:** Creating summaries of instructional materials, such as those found on WikiHow, to provide quick overviews of procedures.

- **Email Subject Generation:** Generating subject lines for emails based on their content, enhancing email management and organization.

Implementation Summary

1. **Loading the Pretrained Pegasus Model and Tokenizer**

   Using Hugging Face's transformers library, to load the pretrained google/pegasus-cnn_dailymail model along with its tokenizer. This model is specifically fine-tuned for abstractive summarization tasks. The model is set to evaluation mode and moved to the GPU to leverage faster computation.

2. **Preparing the Dataset**

   The SAMSum dataset, comprising approximately 16,000 messenger-like conversations paired with summaries, is downloaded manually and loaded using the Hugging Face datasets library. This dataset is particularly suited for training and evaluating dialogue summarization models

3. **Tokenizing Inputs and Generating Summaries**

   Each conversation in the dataset is tokenized using the Pegasus tokenizer, ensuring that inputs are appropriately truncated or padded to fit the model's expected input size. The tokenized inputs are then passed through the Pegasus model to generate summaries.

4. **Fine Tuning and Evaluating Model Outputs**

   The generated token sequences are used for training the model. To assess the quality of the generated summaries, evaluation metrics such as ROUGE scores are computed by comparing the model outputs against the reference summaries provided in the dataset.

5. **Saving the Fine-Tuned Model and Tokenizer**

   After training and evaluation, the fine-tuned Pegasus model and its tokenizer are saved to disk. This allows for easy reuse of the model for inference or further fine-tuning without retraining from scratch.

Major Findings & Conclusion

- **Pegasus Excels at Abstractive Summarization of Dialogues**

  The pretrained model demonstrates strong performance in generating concise and coherent summaries from conversational text, showcasing its effectiveness on dialogue-based datasets like SAMSum.

- **High-Quality Summaries with Minimal Fine-Tuning**

  Even without extensive fine-tuning, the model produces summaries with high ROUGE scores, indicating strong out-of-the-box generalization to dialogue summarization tasks.

- **Reusable Pipeline for Summarization Tasks**

  The saved model and tokenizer allow for rapid reuse in production or experimentation, supporting scalable deployment for summarizing varied textual inputs.

- **Evaluation Metrics Provide Insight into Model Performance**

  Using ROUGE as an evaluation metric offers a quantitative measure of summary quality, helping identify the model's strengths and areas for improvement in different conversational contexts.

References:

1. https://research.google/blog/pegasus-a-state-of-the-art-model-for-abstractive-text-summarization/
2. https://ritvik19.medium.com/papers-explained-162-pegasus-1cb16f572553