Enhancing Summarization of Specific Domain Long Documents through Fine-Tuning Longformer on a Focused Dataset

Analytics Capstone Project

Dmitrii Bakhitov

Instructor: Dr. Christelle Scharff

Pace University, Seidenberg School of CSIS

Abstract

This study aimed to develop an efficient summarization model for machine learning articles by fine-tuning a Longformer model on a focused subset of the arXiv scientific dataset. The performance of the fine-tuned model was compared to state-of-the-art models trained on the full arXiv dataset, demonstrating improved summarization results. This highlights the importance of training natural language processing models on focused and relevant datasets for specific tasks.

Research Question

How to build a deep learning-based summarization model that can generate accurate and consistent summaries of machine learning articles?

Related Work

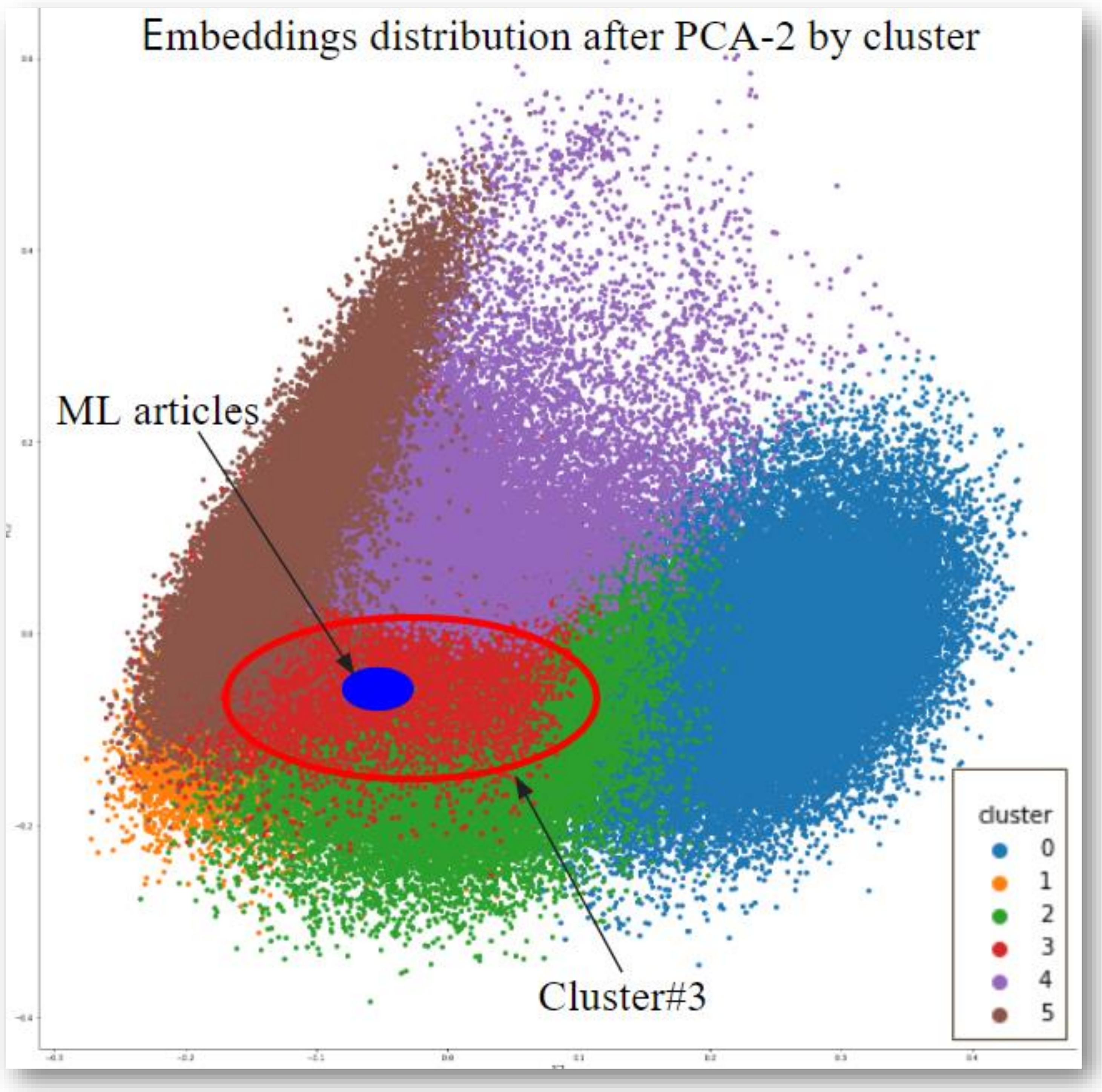
Text summarization has gained significant attention in recent years, with three primary approaches described in [1]: extraction-based, abstraction-based, and hybrid methods. Deep learning-based abstractive summarization techniques have become increasingly popular [2]. In [5], a transformer-based model, Longformer, was proposed, effectively handling long documents using a sliding window attention mechanism and global attention component. Recent works [6, 7] propose hybrid methodologies for summarizing long documents, combining local attention, content selection mechanisms, and global and local tokens.

Dataset

The 'Scientific papers' dataset [8] contains 215,913 scientific papers and summaries in areas of computer science, math, physics, quantitative biology, and finance obtained from the ArXiv repository. This dataset has been used for training summarization models in [3, 4, 5, 6, 7]. A focused subset of this dataset was generated for the present study.

Methodology

To create the subset, 2,077 machine learning-related articles were identified within the 'Scientific papers'. To expand the subset, sentence embeddings were extracted from all article summaries using SciBERT [9]. These embeddings were clustered into six groups using K-means clustering, the clusters number determined using the elbow method. Cluster #3, containing 30,280 instances, was selected as most closely related to ML topic based on cosine similarity.



Model Fine-tuning

The base version of the Longformer (LED) model [5] was fine-tuned on the selected subset for five epochs. The input size was reduced to 7,168 tokens due to GPU memory limitations. The fine-tuning process took over 150 hours on an Nvidia RTX 3070.

Evaluation

The performance of the fine-tuned model was evaluated using the ROUGE metrics [10], which measure the overlap between the ground truth summary and the generated summary in terms of unigrams (ROUGE-1), bigrams (ROUGE-2), and the longest contiguous common sequence (ROUGE-L).

Results

The fine-tuned Longformer model, despite input size limitation, demonstrated superior performance when compared to state-of-the-art models trained on the entire arXiv dataset. The ROUGE scores for the fine-tuned model were consistently higher across all three metrics (ROUGE-1, ROUGE-2, and ROUGE-L) compared to baseline models. This indicates that the model was successful in generating summaries that closely resembled the ground truth summaries in terms of content and structure. The improved performance can be attributed to the focused and relevant dataset used for fine-tuning, emphasizing the importance of domain-specific training data in NLP tasks.

MODEL	ROUGE-1	ROUGE-2	ROUGE-L
BART-4096-arxiv	0.412	0.161	0.234
Pegasus-arxiv	0.398	0.142	0.222
BigBird-Pegasus-large-arxiv	0.402	0.148	0.231
LED-large-16384-arxiv	0.436	**0.168**	0.245
LED-large-10240-arxiv	0.433	0.161	0.239
LED-base-7168-fine-tuned(OUR)	**0.442**	**0.168**	**0.248**

Conclusion & Future Work

This study showcases the effectiveness of fine-tuning a Longformer model on a focused subset of the arXiv dataset for enhancing the summarization of machine learning articles. The results highlight the importance of leveraging domain-specific training data and the benefits of tailoring models to particular subjects or topics. The improved performance achieved by the fine-tuned model suggests that this approach can lead to more accurate and consistent summaries for machine learning articles. Future work will explore experiments with different clustering techniques and subset selection, transformers, and input size.

References

1) Allahyari, Mehdi, et al. "Text summarization techniques: a brief survey." arXiv preprint arXiv:1707.02268 (2017).

2) Zhang, Mengli, et al. "A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning." Computational Intelligence and Neuroscience 2022 (2022).

3) Koh, Huan Yee, et al. "An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics." ACM computing surveys 55.8 (2022): 1-35.

4) Dong, Yue, Andrei Mircea, and Jackie CK Cheung. "Discourse-aware unsupervised summarization of long scientific documents." arXiv preprint arXiv:2005.00513 (2020).

5) Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).

6) Manakul, Potsawee, and Mark JF Gales. "Long-span summarization via local attention and content selection." arXiv preprint arXiv:2105.03801 (2021).

7) Zaheer, Manzil, et al. "Big bird: Transformers for longer sequences." Advances in neural information processing systems 33 (2020): 17283-17297.

8) Cohan, Arman, et al. "A discourse-aware attention model for abstractive summarization of long documents." arXiv preprint arXiv:1804.05685 (2018).

9) Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A pretrained language model for scientific text." arXiv preprint arXiv:1903.10676 (2019).

10) Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.