Redefining Summaries: GANs and Transformers in Abstractive Text Summarization

Farzan Mirza
Drexel University
fm474@drexel.edu

Nakul Narang
Drexel University
nn474@drexel.edu

Abstract

This research proposes an innovative approach to abstractive text summarization utilizing a generative adversarial network (GAN) framework. Targeting the creation of summaries that are indistinguishable from human-generated content, this study uses a Generator that generates the summary and a Discriminator that classifies the generated summary as being machine-generated or human generated. What separated this GAN approach from others is the use of a pre-trained Facebook BART model that has a bi-directional encoder like BERT and an autoregressive GPT-like decoder. By aiming to surpass existing ROUGE score benchmarks, this project not only contributes to theoretical advancements in natural language understanding and generation but also explores practical applications in diverse fields such as news aggregation and educational tools.

1 Introduction

Abstractive text summarization endeavors to distill the essence of lengthy texts into concise, informative summaries. This form of summarization not only condenses content but also creatively rephrases and restructures the original text, introducing new phrases and constructs that were not present in the source document. Extensive research has been undertaken in this domain (Nallapati et al., 2016; See, Liu, and Manning, 2017; Paulus, Xiong, and Socher, 2017), showcasing significant advancements. However, several challenges persist in abstractive summarization: (i) Neural sequence-to-sequence models often yield overly simplistic and generic summaries dominated by high-frequency phrases; (ii) These models sometimes struggle to maintain grammaticality and readability in the generated text; (iii) Traditional training approaches, predominantly based on maximum-likelihood estimation (MLE), face limitations. They align poorly with real-world evaluative metrics and introduce discrepancies between training and testing phases due to their reliance on ground truth inputs during training, which shifts to previously generated outputs during testing. This shift can lead to the accumulation of errors, a phenomenon known as exposure bias.

Addressing these challenges, this paper proposes an innovative adversarial framework to refine the training of abstractive summarization models. We introduce a dual-model system, comprising a generative model (G) and a discriminative model (D). The generative model processes the input text to produce summaries, while its training is enhanced through reinforcement learning techniques, specifically policy gradients. This method circumvents the pitfalls of exposure bias and the limitations of non-differentiable task metrics. Meanwhile, the discriminative model functions as a text classifier, trained to differentiate between machine-generated and human-like summaries. This adversarial interaction promotes the generation of plausible, high-quality abstrative summaries, pushing the boundaries of current technology in natural language processing.

2 Background and Related Work

Abstractive text summarization aims to condense extensive texts into succinct, informative summaries that capture the essence of the content through new phrases and sentences. Over the past decades, this task has evolved significantly with the advancement of deep learning techniques, shifting from traditional extractive methods to more sophisticated abstractive approaches. Early work in this field by Nallapati et al. (2016) laid the groundwork by employing sequence-to-sequence models to tackle the challenge of generating coherent and concise summaries directly from source texts Nallapati et al. (2016).

Despite the advancements, several inherent challenges persist in abstractive summarization.

Neural sequence-to-sequence models often yield summaries that are either too generic or overly reliant on high-frequency phrases, thus compromising uniqueness and informativeness. These models typically employ maximum-likelihood estimation (MLE) for training, which inadvertently leads to a discrepancy between training objectives and actual evaluation metrics, as well as the phenomenon known as exposure bias during testing phases (See, Liu, and Manning, 2017) See et al. (2017).

The introduction of the pointer-generator network by See, Liu, and Manning (2017) marked a significant improvement in handling these challenges. This model enhances the sequence-to-sequence framework by integrating a copying mechanism that allows for the inclusion of out-of-vocabulary words directly from the source text, which helps mitigate issues of repetitiveness and phrase commonality in the generated summaries See et al. (2017).

Furthering this line of inquiry, Paulus, Xiong, and Socher (2017) integrated reinforcement learning with abstractive summarization to optimize the quality of summaries based on actual performance metrics rather than just likelihood maximization. Their approach, which directly aligns the training process with the evaluation metrics, addresses key shortcomings of the traditional training methods by reducing the impact of exposure bias, thus improving both the relevance and the fluency of the generated summaries Paulus et al. (2017).

Our research builds upon these foundational studies and adopts an adversarial training framework, inspired by the work of Goodfellow et al. (2014) on generative adversarial networks (GANs). This methodology introduces a competitive dynamic between a generative model that creates summaries and a discriminative model that judges their quality, enhancing the generative capabilities of the summarization model to produce outputs that are not only grammatically correct and coherent but also indistinguishable from human-written text Goodfellow et al. (2014).

Most recently, the BART model introduced by Lewis et al. (2020) has significantly advanced the performance on a broad range of summarization tasks. BART employs a denoising autoencoder framework for pre-training sequence-to-sequence models, which has proven effective in enhancing the quality of generated text across different domains, setting a new benchmark for subsequent models in natural language processing Lewis et al. (2019).

Incorporating insights from these studies, our research aims to extend the adversarial framework to further address the limitations noted in prior works, focusing on generating coherent, grammatically sound, and human-like summaries.

3 Methodology

In this study, we implement a generative adversarial network (GAN) comprising of two primary components: the generative model (Generator G) and the discriminative model (Discriminator D). The Generator G is built using Facebook BART long model Lewis et al. (2019). This model processes input texts—which are tokenized, encoded and embedded using tokenizer function provided by BART — into concise summaries. The BART model used has already been pre-trained on CNN/DailyMail corpus data. The generative model is uniquely trained using policy gradients to optimize the quality of summaries based on feedback from the Discriminator D (treated as reward), effectively bypassing exposure bias.

Conversely, the Discriminator D utilizes a Convolutional Neural Network (CNN) architecture that classifies texts as either human-generated or machine-generated. This model leverages BART tokenizer to convert text into dense, meaningful representations. It features multiple convolutional layers with varying filter sizes to extract a broad spectrum of textual features, which are then condensed through max pooling. The training of the Discriminator focuses on enhancing its ability to discern between human-written summaries and those produced by the Generator, using cross entropy as the objective function.

The adversarial training of both models is setup wherein the Generator aims to fool the Discriminator into classifying its generated summaries as human-like, thereby refining its output iteratively, while, the Discriminator strives to accurately distinguish between the two types of summaries, enhancing its evaluative accuracy. This dynamic interplay not only improves the Generator's ability to produce plausible summaries but also sharpens the Discriminator's evaluative precision. The generator is treated as a policy and it is updated using the output of the discriminator as a reward. The above idea makes sense as the output of the discriminator is the likelihood of the summary being

either machine generated or not.

Liu et al. (2017) has shown that after getting summaries from G, we can re-train D using the below equation:

$$\min -E_{Y\sim p_{\text{data}}}[\log D_\phi(Y)] - E_{Y\sim G_\theta}[\log(1 - D_\phi(Y))]$$

Our policy gradient update is also followed from the work Liu et al. (2017), which is depicted by the below equation:

$$\begin{aligned}\nabla_\theta J_{pg} =& \frac{1}{T}\sum_{t=1}^{T}\sum_{y_t} R_D^{G_\theta}\left((Y_{1:t-1}, X), y_t\right)\cdot\nabla_\theta\left(G_\theta\left(y_t \mid Y_{1:t-1}, X\right)\right)\\ =& \frac{1}{T}\sum_{t=1}^{T}\mathbf{E}_{y_t\in G_\theta}\left[R_D^{G_\theta}\left((Y_{1:t-1}, X), y_t\right)\nabla_\theta \log p\left(y_t \mid Y_{1:t-1}, X\right)\right]\end{aligned}$$

For our experiments, we employ the CNN/DailyMail dataset, chosen for its extensive collection of news stories and accompanying summaries, providing a rich training ground for our models. The dataset is divided into training, validation, and testing sets, ensuring a comprehensive evaluation framework. The effectiveness of the generated summaries is quantified using ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L), which measure the overlap of unigrams, bigrams, and the longest common subsequence, respectively, between the machine-generated summaries and reference summaries. Additionally, we plan to incorporate human evaluations to assess the summaries' coherence, relevance, and readability.

This methodology ensures a robust framework for advancing the state of abstractive text summarization, aiming to produce summaries that are not only concise and informative but also indistinguishable from those written by humans.

4 Analysis and Conclusions

4.1 Analysis

Our model was trained 16 samples from the dataset and our test set comprised of 2 samples. Due to computation limitations we had to go with a diminished version of the dataset. In our analysis we have compared the performance of our model against ABS Nallapati et al. (2016), PGC See et al. (2017), DeepRL Paulus et al. (2017) and Bi-LSTM Liu et al. (2017). The analysis of ROUGE scores from Table 1 indicates that our model performed better than its peers in all metrics. In ROUGE-1 our model beat the benchmark by 3 points. Similarly for ROUGE-2 and ROUGE-L our model beat the corresponding benchmarks by approximately 16 points and 1 point.

Methods	ROUGE-1	ROUGE-2	ROUGE-L
ABS	35.46	13.30	32.65
PGC	39.53	17.28	36.38
DeepRL	39.87	15.82	36.90
Pretrain	38.82	16.81	35.71
Bi-LSTM	39.92	17.65	36.71
Ours	42.48	33.98	37.22

Table 1: Results of the experiment.

The summaries generated were human readable and were semantically and grammatically coherent. An example of the evolution of a summary of an article with the epochs is given in Appendix A.

4.2 Conclusions

The overwhelming performance of our model emphasizes the importance of adversarial training paradigm in abstract summary generation. The summaries generated were coherent with the article and were found to be of acceptable standards for human reading.

5 Future Work

In future work, we aim to expand the scope and capabilities of our abstractive text summarization model in several key areas. Firstly, we aim on training our model on the entire dataset to get the full extent of its capabilities and performance. Secondly, incorporating multimodal data, such as images and videos, could enhance the contextuality and richness of summaries, addressing the need for summarization across diverse media types. Additionally, expanding the diversity of languages and datasets—including scientific articles, novels, and legal documents—will help evaluate the model's robustness and adaptability to various domains. Exploring more sophisticated discriminative models, could refine the discriminator's evaluative accuracy. Lastly, real-world deployment and testing of the model in practical, user-centric scenarios will provide invaluable insights into its effectiveness and user satisfaction, paving the way for broader application and impact in automated text summarization.

References

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative ad-

versarial nets. In *Advances in Neural Information Processing Systems (NIPS)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2017. Generative adversarial network for abstractive text summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstrative summarization. *arXiv preprint arXiv:1705.04304*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

A Appendix A

In this appendix we have included the summaries generated by the GAN at each epoch:

Original Summary: Bomb victims waiting for presidential visit . Blast went off 15 minutes before president's arrival . Algeria faces Islamic insurgency . Al Qaeda-affiliated group claimed July attacks .

Epoch 1: Bomb victims waiting for presidential visit . Blast went off 15 minutes before president Abdel-Aziz Bouteflika's arrival in Batna . Al Qaeda-affiliated group claimed July attacks in Algiers .

Epoch 2: Bomb victims waiting for presidential visit . President Abdel-Aziz Bouteflika says terrorist acts have nothing to do with Islam . In July, 33 people killed in apparent suicide bombings in Algiers .

Epoch 3: Bomb victims waiting for presidential visit . Device exploded 20 meters from mosque in Batna, east of capital of Algiers . Blast went off 15 minutes before president's arrival . President Abdel-Aziz Bouteflika says terrorist acts have nothing in common with noble values of Islam .