Abstractive Scientific Text Summarization using Generative Adversarial Networks

Maria Dobko
Ukrainian Catholic University
Faculty of Applied Sciences
Lviv, Ukraine
dobko_m@ucu.edu.ua

Oleksandr Zaytsev
Ukrainian Catholic University
Faculty of Applied Sciences
Lviv, Ukraine
oleks@ucu.edu.ua

Yuriy Pryima
Ukrainian Catholic University
Faculty of Applied Sciences
Lviv, Ukraine
y.pryima@ucu.edu.ua

Abstract

Generative adversarial networks (GAN) have shown a lot of success in image generation. However until recent years they were considered inapplicable to the discrete problems of natural language processing (NLP). The latest papers introduce novel approaches to overcoming these issues by combining GANs with reinforcement learning (RL) models and lay the foundation for the whole new field of research of adversarial language processing.

In our research we will apply GANs to the task of scientific text summarization and try to improve the results of recent state-of-the-art approaches.

Keywords text summarization, NLP, GAN, reinforcement learning

1 Introduction

Recent studies have shown that neural networks can be used in solving NLP problems. However, models that were mostly considered for this task were Convolutional Neural Networks and Recurrent Neural Networks.

Applying generative adversarial networks to the problems of NLP is a considered to be a complicated task because GANs are only defined for real-valued data, and all NLP is based on discrete values like words, characters, or bytes.

> For example, if you output an image with a pixel value of 1.0, you can change that pixel value to 1.0001 on the next step. If you output the word "penguin", you can't change that to "penguin + .001" on the next step, because there is no such word as "penguin + .001". You have to go all the way from "penguin" to "ostrich".[1]

However, in their latest paper Fedus, Goodfellow, and Dai[FGD18] overcome this problem by using Reinforcement Learning (RL) to train the generator while the discriminator is still trained via maximum likelihood and stochastic gradient descent, and use it to fill the gaps in the text.

However, this approach has not been used for abstract text summarization of scientific papers, yet.

2 Problem statement

The main issue of using generative adversarial networks in Natural Language Processing is connected with the fact that text data is discrete. Why does it make a difference? Generative Adversarial Network consists of two models: first is generative, which tries to fool with synthetic data the second model - discriminative, that distinguishes the difference between real and generated data. As it is stated in [FGD18], GANs have had a lot of success in producing more realistic images than other approaches but they have only seen limited use for text sequences. GAN?s are widely used on images, where generating a new sample requires only to change the density of specific pixels. While working with text it becomes more difficult to tell the generator about how to change the input, thus, according to [LLY$^+$17] it is a good choice to build the generator as an agent of reinforcement learning, which takes the raw text as input and predicts the abstractive summarization.

2.1 Research Question/Hypothesis

The main hypothesis is whether Generative Adversarial Network can work better than other methods in the presented task: abstract text summarization of scientific paper.

3 Background and significance

Allahyari et al.[APA$^+$17] make a survey of the most successful text summarization techniques as of July 2017.

This September Li et al. [LCB17] described their submission to the sentiment analysis sub-task of ?Build It, Break It: The Language Edition (BIBI)? where they successfully apply generative approach to the problem of sentiment analysis.

In their paper *Generative Adversarial Network for Abstractive Text Summarization* Liu et al.[LLY$^+$17] built an adversarial model that achieved competitive ROUGE scores with the state-of-the-art methods on CNN/Daily Mail dataset. They compare the performance of their approach with three methods, including the abstractive model, the pointer-generator coverage networks, and the abstractive deep reinforced model.

4 Research design and methods

We will write a script for downloading a given number of papers from arXiv. Those papers will be filtered by their category and publication date.

[1] Ian Goodfellow's answer to the related question on Reddit: https://www.reddit.com/r/MachineLearning/comments/40ldq6/generative_adversarial_networks_for_text/

4.1 Data collection

arXiv provides a RESTful API[2] that allows us to search for papers from a specific category and inside a specific time range. The results are returned as an HTML page which can be easily parsed. By making requests to arXiv and parsing the response we acquire all the necessary information about the paper (paper id, title, authors, date, subjects, and abstract)[3]. Then we use the collected list of paper ids to download the PDF files, we extract text from those files and store it in a table together with other variables acquired from arXiv.

4.2 Timeframes

February 28	deadline for data collection
March 7	deadline for data analysis and corpus training
March 21	trying different architectures, comparing results
April 30	final paper submission

5 Strength and weakness of the study

As our main strength we see a possibility to test different GAN architectures in order to find the one, that will work good for text summarization. As long as, this topic is pretty hot and in the same time poorly studied, we have a freedom to choose which dataset to use for training, thus it is also a part of research. However, there are high risks that this approach may not give good results at all, because there are still lots of issues about usage of neural networks with language data. The other possible weakness is a difficulty to compare the results with other papers, as there are not a lot of researches concerning this or relative subject. We are also currently looking for supervisors who might be interested in the following topic, so we could have a mentorship during the research.

References

[APA+17] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. Text summarization techniques: A brief survey, 2017.

[FGD18] William Fedus, Ian Goodfellow, and Andrew M. Dai. Maskgan: Better text generation via filling in the gaps, 2018.

[LCB17] Yitong Li, Trevor Cohn, and Timothy Baldwin. Bibi system description: Building with cnns and breaking with deep reinforcement learning, 2017.

[LLY+17] Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. Generative adversarial network for abstractive text summarization, 2017.

[2] https://arxiv.org/help/api/index

[3] Our first attempts of data scrapping from arXiv: https://github.com/MachineLearningUCU/arXiv-parsing