# 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 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. We will take the whole text of a paper and produce a shorter text that contains a concise summary of the research. Sophisticated structure of scientific texts and the availability of many additional sources of contextual information (such as the referenced papers and the topics of other papers written by the authors). We will compare the performance of several discrete GANs that performed well on the similar problems of text generation, and try to improve the results of recent state-of-the-art approaches to scientific text summarization.

***Keywords*** text summarization, NLP, GAN, reinforcement learning

## 1 Introduction

Recent studies have shown that neural networks can be used for 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]

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

However, in their latest paper Fedus, Goodfellow, and Dai[FGD18] overcome this problem by using reinforcement learning 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 hypothesis is that generative adversarial networks with reinforcement learning based generator, when applied to the problem of abstractive scientific text summarization, can provide better results than recent state-of-the-art approaches.

## 3 Related work

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.

In contrast Zhang et al.[GFC+17] don't use reinforcement learning but rather introduce TextGAN with an LSTM-based generator and kernelized discrepancy metric.

## 4 Research design and methods

We will be applying existing discrete GANs to the data collected from arXiv. One of the lates successful models for scientific text summarization will be selected as our baseline.

### 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

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

be easily parsed. By making requests to arXiv and parsing the response we acquire all necessary information about the paper (id, title, authors, date, subjects, abstract etc.)[3]. Then we use the collected list of paper ids to download PDF files, extract text from those files and store it in a table together with other variables acquired from arXiv.

### 4.2 Timeframes

***Deliverables for the*** $1^{st}$ ***evaluation***

- Dataset of papers collected from arXiv
- Results of feature extraction
- Implementations of the baseline state-of-the-art model and its application to our dataset

***Deliverables for the*** $2^{nd}$ ***(final) evaluation***

- Implementation of several discrete GAN models
- Evaluation of the created models on our dataset
- Paper describing the results of our research

## 5 Strength and weakness of the study

GANs have proved to be the most successful when it comes to generative images, but their application to the problems of the text generation is not well studied. So we expect our research to introduce novel approaches and original ideas that may advance the field of natural language processing. 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.

[GFC+17] Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation, 2017.

[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.


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