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

The aim of this research is to test whether it is possible to use generative adversarial networks to solve the task of abstract text summarization. Achieved results will help better understand how neural networks can be used for NLP. Furthermore, using the dataset of scientific papers collected from arXiv and other similar resources we can test how good our model works in summarizing text. For training we will create our own corpus from scientific papers, as labels we will choose ?Abstract? part from each of them.

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. Only during the last years papers about experiments of using GAN?s in NLP were published. For example, authors of [**?**] use generative approach for sentiment analysis, in [**?**] a new architecture of GAN is proposed for filling the gaps in texts, some experiments using generative models for text summarization were shortly described in [**?**]. 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 [**?**], 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 [**?**] 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

Until very recently it was considered impossible to apply GANs to the problems of NLP[1].

Allahyari et al.[**?**] make a survey of the most successful text summarization techniques as of July 2017.

4 Research design and methods

4.1 Overview

4.2 Population and Study Sample

We will write a script that will download a given number of papers from arXiv.

4.3 Sample Size and Selection of Sample

4.4 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.

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

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

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

4.5 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.