Computer Science Research Ideas Generation Using Neural Networks

Ashwag Maghraby[1†] and Joanna Assaeed[2††],
aomaghraby@uqu.edu.sa s44380113@st.uqu.edu.sa
Umm Al-Qura University, College of Computer and Information Systems ,Makkah, Saudi Arabia

Summary
The number of published journals, conferences, and research papers in computer science is increasing rapidly, which has led to a challenge in coming up with new and unique ideas for research. To alleviate the issue, this paper uses artificial neural networks (ANNs) to generate new computer science research ideas. It does so by using a dataset collected from IEEE published journals and conferences to train an ANN model. The results reveal that the model has a 14% success rate in generating usable ideas. The outcome of this paper has implications for helping both new and experienced researchers come up with novel research topics.

Keywords:
Neural Networks; Artificial Intelligence; Computer Science Research Ideas; Topic Generation; GPT2

1. Introduction

The rate of computer science publications is growing rapidly, with over four million articles and conference papers published in the past 20 years in IEEE [1] alone. Even within a specific field and considering all the publishing libraries, a researcher must invest a considerable amount of time to keep up with all the new developments in computer science. Finding and discovering new research topics and all the relevant information can be quite challenging for both new and veteran researchers.
Different researchers have used a variety of machine-learning methods to generate hypotheses and ideas in different fields. Machine learning is a subfield of artificial intelligence that aims to use computational algorithms to replicate the ways in which humans learn [2,3]. Sang et al. [4] used the traditional activity-based cost (ABC) model and co-training algorithm (a semi-supervised learning algorithm used when the labeled data is relatively small compared to the unlabeled data) to generate a biomedical hypothesis. Ultimately, the authors proved their method was better than using cooccurrence and grammar-engineering approaches. Friederich et al. [5] used a data-driven workflow to generate hypotheses in the field of natural science. To demonstrate the abilities of this approach, they successfully applie it to rediscover known knowledge in chemistry. However, the automated workflow only functioned in applications that can be represented as graphs.

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.19

In contrast, Gonsalves [6] combined a machine-learning and the backbone of deep learning algorithms known as artificial neural networks (ANNs), a field that focuses on imitating human brain computation [7,8,9]. The author used several open-source projects and custom python codes to train a neural network to generate new ideas. The InventerBot project used custom-written python code for preprocessing and post-processing and open-source projects to train the neural network, as shown in Fig.1. However, the results showed that about 90% of the generated ideas were poorly written.

The main purpose of the present study is to generate new computer science research topics using neural networks to support research progress in computer science. It also aims to disprove the null hypothesis that generating computer science topics using neural networks will not lead to new research topics. The remainder of this paper comprises the dataset and the training method in Section 2, the results and discussion in Section 3, and the conclusion and future work in Section 4.

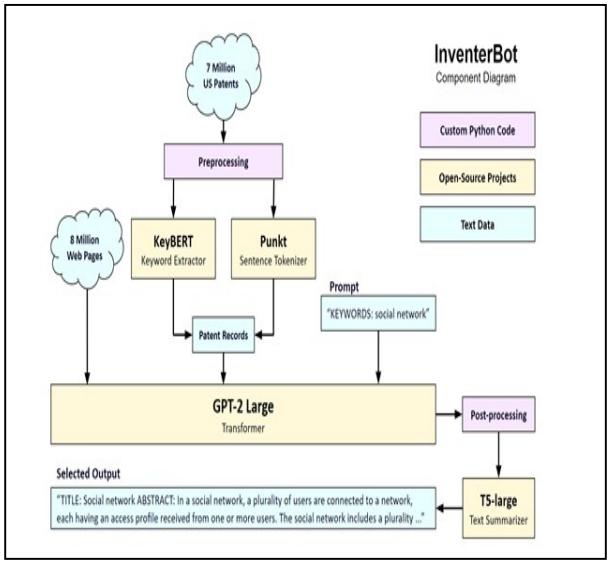

Fig. 1 NventerBot Component Diagram [6]

2. Computer Science Ideas Generator

Due to the ever-increasing number of research publications in the field of computer science, forming new and unique research ideas is becoming more and more challenging.

2.1 Dataset

2.1.1 Description

To train the neural network model Generative Pretrained Transformer (GPT2) [10], the abstracts, titles, and publication dates of 500 published journals and conference papers were collected from IEEE Application Programming Interface (API) [11]. The data collected was primary data due to the unavailability of a similar dataset. IEEE API was chosen as a data source due to the organization's high publication standards and the availability and usability of the API. To begin the collection process, an IEEE API key for "metadata search" was acquired. Afterward, a search query was constructed using the start year, end year, and content type as filters. The start year and end year were set to "2016" and "2021," respectively, to limit the publication date of the results to the past five years and to test the results against published research. Content type was set to "conferences" and "journals" to constrain the results to only include what is considered a research paper. The search query was then used in a python code to collect 500 journals or articles and store them in a JSON file. The data collection process lasted three days due to the 200-call-per-day limit that IEEE sets. Table 1 shows some of the fields the JSON file returned from the API request.

2.1.2 Dataset Value

The data were collected from IEEE, one of the most known and respectable organizations in computer science and engineering research. For this reason, the data are beneficial to anyone looking to replicate this research result or to further the work on this topic. They could also be applied to other research topics requiring metadata from IEEE.

2.1.3 Pre-Processing

Following Gonsalves' [6] methodology, the data must be preprocessed before passing to the neural network model. The JSON file obtained from IEEE had several extraneous fields and lacked the organization necessary for it to be usable in training the neural network model. KeyBERT [12], a keyword extractor, and Punkt [13], a sentence tokenizer were used to preprocess the data. First, KeyBERT was applied to every paper's title to extract 1–3 relevant keywords. For example, using KeyBERT on the title "Using Model Checking to Detect Simultaneous Masking in Medical Alarms" would yield the keywords "masking," "medical," and "alarms." Punkt is then applied to each of the paper's abstracts to generate a summary by creating a list with each abstract sentence as an element and then picking the first two sentences as a summary. In the end, a text file is created that includes each processed paper keyword, title, and the short abstract (see Fig. 2).

Table 1: Samples of some of the json data fields

Abstract_Url	Access Type	Article Number	Content Type	Publication_Year	Title
https://ieeexplore.ieee.org /document/5765911/	LOCKED	5765911	Magazines	June 2016	Integrated Systems in the More-Than-Moore Era: Designing Low-Cost Energy-Efficient Systems Using Heterogeneous Components
https://ieeexplore.ieee.org /document/5936050/"	LOCKED	5936050	Magazines	Dec. 2016	Handling Nondeterminism in Logic Simulation so That Your Waveform Can Be Trusted Again
https://ieeexplore.ieee.org /document/6335467/	LOCKED	6335467	Journals	April 2016	Sparsity-Induced Similarity Measure and Its Applications
https://ieeexplore.ieee.org	LOCKED	6357187	Journals	1 Aug. 2017	Wait-Free Programming for

Abstract_Url	Access Type	Article Number	Content Type	Publication_Year	Title
/document/6357187/					General Purpose Computations

```
KEYWORDS: ('ghz integrated cmos', 0.6296)
TITLE: A 5.8 GHz Integrated CMOS Dedicated Short Range Communication
Transceiver for the Korea/Japan Electronic Toll Collection System
ABSTRACT: In this paper, a RF front-end of the 5.8 GHz integrated
CMOS dedicated short range communication (DSRC) transceiver for the
Korea/Japan electronic toll collection system is presented. The
receiver uses low-IF conversion architecture for high sensitivity and
low-power consumption while the transmitter uses direct up-conversion
architecture for its simple structure and reliability.
***
```

Fig. 2 Sample of the pre-processing results

2.1.4 Training Methodology

This paper chose Gonsalves' training method due to similarities between the author's project and the objective of the present paper. This paper chose Gonsalves' training method due to similarities between the author's project and the objective of the present paper.

Gonsalves' use of open-source projects also granted flexibility and accessibility to the method. Therefore, the GPT2 was trained using the IEEE preprocessed datasets on Google Colab. The model with 124 million parameters was chosen and trained for 500 steps for the GPT2 model configurations.

3. Results and Discussion

After training the GPT2 model, it was prompted to use different keywords to generate new research titles and abstracts. The generated abstracts were a bit lengthy; therefore, Google's Text-to-Text Transfer Transformer (T5) [14] was used to create briefer abstracts. Table 2 contains samples of the results the model generated. To evaluate the results, six different keywords were used to generate 100 new research ideas. These ideas were then compared with existing publications to determine whether any of them would be viable research topics. The six keywords were chosen due to their high usage in published research and included "data mining," "security," "machine learning," "graphics," "image processing," and "IoT." Out of 100 generated research titles, only 14 resembled existing literature, 4 were repeated titles, 5 were generic titles not suitable as a research topic, and 77 were not found in the literature. Ultimately, the model presented in this paper had a 14% success rate of generating applicable research ideas.

4. Conclusion

This research aimed to generate new computer science research ideas. Based on the data collected from IEEE and the GPT2 training model, it is possible to generate new computer science research ideas. The results demonstrate that the research ideas the model generated had a 14% chance of being viable research topics. This research also illustrates the capabilities of the GPT2 model and natural language process (NLP) models. In terms of limitations, the research is limited by the size of the dataset. Future studies should use a larger dataset and larger NLP models such as GPT3 and switch transformers.

Table 2: Sample of gpt2 research ideas generation

keyword	Title	Abstract
Data mining	Data Mining: A High-Frequency Oligopolistic Potential in Vehicular Ad Hoc Networks	The potential of vehicular oligopolistic potential is evident in the network architectures and their interference policies.
Security	Secure Key Exchange and Its Implementation in Smart Grid	The secure key (skey) of a wireless network is determined by the number of traffic sources. This paper analyzes the problem of key sharing in wireless networks with two traffic types.
Machine Learning	Machine Learning to Read from a Memory List via Forwarding and Echoing	Memory technology has emerged as the future of computing. A number of machine learning models have been focusing on memory management.
Graphics	A Game-Theoretic Resource Allocation Approach to Graphics	New types of resources, opportunistic or not, may arise from mobile devices. These resources may be used in personalized ways to a specific user.
Image Processing	Image Reconstruction From Magnetic Fields	Direct image reconstruction is a technique for image reconstruction from magnetic fields. It uses the processor images of a humanoid model to determine its physical structure
IoT	A Routing Protocol for IoT Device-to-Device Communications	This paper proposes a new routing protocol for device-to-device (d2d) communications. It

keyword	Title	Abstract
		provides routing visibility into devices configured to receive iot-connected devices.

References

[1] IBM Cloud Education: *What are Neural Networks?*. Accessible: https://www.ibm.com/cloud/learn/neural-networks

[2] Woolf, M.,: *GPT2 Simple Source Code*. In: Github. (2021) Accessible: https://github.com/minimaxir/gpt-2-simple

[3] IEEE API Main Page. Accessible: https://developer.ieee.org/

[4] Grootendorst, M.,: *KeyBERT Source Code.* In: Github. (2021) Accessible: https://github.com/MaartenGr/KeyBERT

[5] Bird, S., Loper, E., Nothman, J., Darcet, A.,: *Punkt Documentation Page*. Sphinx and NLTK Theme. (2021) Accessible: nltk.tokenize.punkt

[6] Roberts, A.,: *Exploring Transfer Learning with T5: The Text-to-Text Transfer Transformer*. In: Google AI blog. (2020) Accessible: https://ai.googleblog.com/2020/02/exploring-transfer-learning-witht5.html

[7] IEEE Home Page. Accessible: https://ieeexplore.ieee.org/Xplore/home.jsp

[8] Naqa, I., Murphy, M.,: *Machine learning in radiation oncology: Theory and applications*.1st ed. Springer, (2015)

[9] IBM Cloud Education: *What is machine learning?*. Accessible: https://www.ibm.com/cloud/learn/machine-learning

[10] Sang, S., Yang, Z., Li, Z., Lin, H.,: *Supervised learning based hypothesis generation from biomedical literature*. BioMed Research International, pp. 1–12. (2015)

[11] Friederich, P., Krenn, M., Tamblyn, I., Aspuru-Guzik, A.,: *Scientific intuition inspired by machine learning-generated hypotheses*. Machine Learning: Science and Technology, vol. 2, pp. 025–027. (2020)

[12] Gonsalves, R.,: *InventorBot: Using AI to Generate New Ideas in Any Field*. In: Geek Culture. (2021) Accessible: https://medium.com/geekculture/inventorbotusing-ai-to-generate-new-ideas-in-any-field-9345f9042df

[13] Georgevici, A., Terblanche, M.,: *Neural networks and deep learning: A brief introduction*. Intensive Care Medicine, vol. 45, pp. 712–714. (2019)

[14] Roy, A.,: *Artificial neural networks: A science in trouble*. SIGKDD Explorer Newsletter, vol. 1, pp. 33–38. (2000)

Ashwag Maghraby received the B.S. (2002) and M.S. (2007) degrees in computer science from King Abdulaziz University, Jeddah University and the Ph.D. degree in Intelligent software engineering, The Centre for Intelligent Systems and their Applications (CISA)University of Edinburgh , 2013. Since 2013, she has been an Assistant Professor with the Department of Computer Science, Umm Al-Qura University, and Mecca, Saudi Arabia. For the last three years her student researches awarded the best college research project in Umm Al-Qura University. Her research interest includes software and intelligent systems engineer which focuses on using NLP and machine learning to improve health care and enhance people's daily lives**.**

Joanna Assaeed[2] received the B.S. degrees in computer science from the University of Umm Al-Qura University, and Mecca, Saudi Arabia. Right now, she studies M.S. in Umm Al-Qura University. Her research interest includes intelligent systems which focuses on using NLP and machine learning.