Effect of Training Large Language Models For KG Generation on their General Language Abilities: The Cost of Specialization?

by

Sharyar Memon, Navid Rezai, Marek Reformat

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Engineering

Department of Electrical and Computer Engineering

University of Alberta

Abstract

This project was focused on understanding the side effects of ”specializing” a large language model such as GPT2-XL on the task of Knolwedge Graph Generation. It was theorized that this kind of ”specialization” may lead to a decrease in the performance of the models in other language modelling and understanding tasks such as multiple choice selection, text entailment, word sense disambiguation and other similar tasks. This comparision was conducted by taking a baseline (non-specialized) GPT2-XL model and comparing its performance with a ”specialized” GPT2-XL model used for Knowledge Graph Generation.

Preface

A preface is required if you need to describe how parts of your thesis were published or co-authored, and what your contributions to these sections were. Also mention if you intend to publish parts of your thesis, or have submitted them for publication. It is also required if ethics approval was needed for any part of the thesis.

Otherwise it is optional.

See the FGSR requirements for examples of how this can look.

To my Dad

For instilling me in the confidence to ask questions and look for answers.

Programming is the closest thing we have to magic. In a fantasy story, with the right runes, a wizard can do anything. With the right codes, a programmer can do anything.

– Alex Shinsel

Contents

1 Introduction **1**

1.1 Knowledge Graphs 1

1.2 Commmon Sense Knowledge Graphs 1

1.3 Large Language Models 2

2 Objective of the project **3**

3 Resources **4**

3.1 Models and Tokenizers 4

3.2 Datasets 5

3.3 Compute Resources 5

4 Methodology **6**

5 Objective **7**

6 Objective **8**

7 Conclusion **9**

References **10**

Appendix A Background Material **12**

List of Tables

List of Figures

A.1 A supporting figure 13

Chapter 1

Introduction

Here is a test reference [13]. As the general human knowledge base grows, so does the need of storing it in a format that is readable and digestible by a computer. As such, knowledge graphs play a very significant role today in how computer store information and digest it. Generally speaking, knowledge graphs are generated by painstaking manual human effort. This has changed with the advent of Large Language Models that can now be used to augment currently existing knowledge graphs [13]

1.1 Knowledge Graphs

There are multiple definitions of knowledge graphs. The widely known one comes from Google's popularization of the word in 2012 [5] where the authors imply that the knowledge that Google contains is accessible via the Google knowledge graph. On the flip side, authors describe knowledge graphs as RDF graphs (a set of RDF triples) [2]. For our purposes, we will describe knowledge graphs as a set of RDF triples that contain a subject, a property and an object.

1.2 Commmon Sense Knowledge Graphs

Commonsense knowledge plays a crucial role today in many machine learning applications including natural language processing and computer vision. Commonsense is often provided via a number of sources dependening on the application. In order to provide a common source that can play multiple roles, CommonSense Knowledge Graphs (CSKGs) were born [4].

1.3 Large Language Models

Language models, in essense, are probablity distributions over sequences of words [6]. They are used for a variety of purposes that range from the simple such as tab completion to sophisticated text generation and reviewing human written translations. With the advent of larger and more sophisticated language models such as GPT-3[1], the scope of usefuless for language models has expanded significantly. One such use is in the augmentation of currently existing commonsense knowledge graphs [13]. This kind of usage requires that large language models (LLMs) such as GPT-2 be trained on the task of knowledge generation. As per West *et al.*, language models fail to express common sense knowledge when prompted in a zero-shot manner. As such, the authors converted the models to COMET-DISTIL models by training them on a knowledge graph. We suspect, such training, while providing additional capabilities for knowledge graph generation, reduces other language modelling capabilities of the trained model.

Chapter 2

Objective of the project

The goal of this project is to understand how training(fine-tuning) a large language model such as GPT2-Xl to predict tails from a head and a relation **??** can affect its performance in other language modelling tasks when compared to its ‘untrained’ version. This will help identify if this kind of training leads to a loss or a gain in capability for a language model. This is done by comparing the performance of the COMET-GPT2-XL[13] model against the ‘naive’ GPT2-XL on a variety of common modelling tasks and metrics. This project is also a study in how to fit large language models when dealing with limited compute resources as often encountered in open-source projects.

Chapter 3

Resources

3.1 Models and Tokenizers

The models used in this project come from two different sources. The GPT2-XL model is based on the OpenAI's GPT2 model [9]. The GPT2-XL model has 48 attention modules as compared to the 12 found in the base GPT2 model. This gives the GPT2-XL model more capability in terms of language modelling tasks but also requires significant more computing resources to train and work with [8]. This translates to 117 million parameters in the GPT2 model vs 1.6 billion parameters in the GPT2-XL model. Given that the COMET-DISTIL models are based on the GPT2-XL model, we opted to utilize the naive GPT2-Xl model as our baseline for comparison. We downloaded the GPT2-Xl model by using the Hugginface model library [3]. As our tokenizer, we used the pre-trained tokenzier also available from the hugginsface model library [3].

The second model we used was the COMET-DISTIL model [13] that was trained on the ATOMIC-10X knowledge graph. The COMET-DISTIL model was downloaded from the author's provided website that can be found on their githhub page [7]. The tokenizer used with this model was also downloaded from the same source. The tokenizers for both models were adjusted to add EOS tokens to ensure that they could work with the datasets that were used for evaluationg the models.

In order to train and test the models, both models had different language modelling heads (dependening on the task) connected to the models. The heads were then trained on the training set of the datasets and then evaulted on the evaulation set of the datasets.

3.2 Datasets

In order to test the performances of the respective models, several different standard datasets were used. This included the widely recognized SuperGLUE dataset [12]. The second benchmark was based on the CommonsenseQA dataset [11] that tests a model's ability to answer a question based on prior knowledge. The SuperGLUE benchmark on the other hand is made up of several different datasets that test a model on a variety of tasks including BoolQ, CB, COPA and MultiRC.

These two benchmarks were choosen based on their wide variety of tasks and the difficulty associated with each of these tasks. In addition, the original glue benchmark was used to test the performance of models to add an additional comparison point to the study.

3.3 Compute Resources

To train and evaluate the performance of the respective models, Lambda Compute Instances at the University of Alberta were utilized. The machines included two Nvidia N100 Gpus with 24GB of memory. Due to the large size of the GPT2-XL Model (5.2 Billion Parameters), the DeepSpeed[10] library from Microsoft was also utilized to help fit the models within the GPU memory.

Chapter 4

Methodology

Chapter 5

Objective

Hello!

Chapter 6

Objective

Hello!

Chapter 7

Conclusion

Referring back to the introduction (Section **??**), we see that cross-references between files are correctly handled when the files are compiled separately, and when the main document is compiled. When the main document is compiled, cross-references are hyperlinked. The values of the cross-references will change between the two compilation scenarios, however. (Each chapter, compiled on its own, becomes “Chapter 1”.)

Caution: For cross-references to work, when files are compiled separately, the referenced file must be compiled at least once before the referring file is compiled.

References

[1] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language Models are Few-Shot Learners," arXiv, Tech. Rep. arXiv:2005.14165, Jul. 2020, arXiv:2005.14165 [cs] type: article. DOI: 10.48550/arXiv.2005.14165. [Online]. Available: http://arxiv.org/abs/2005.14165 (visited on 06/08/2022).

[2] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger, "Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO," en, *Semantic Web*, vol. 9, no. 1, A. Zaveri, D. Kontokostas, S. Hellmann, *et al.*, Eds., pp. 77–129, Nov. 2017, ISSN: 22104968, 15700844. DOI: 10.3233/SW-170275. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-170275 (visited on 06/08/2022).

[3] *Gpt2-xl · Hugging Face*. [Online]. Available: https://huggingface.co/gpt2-xl (visited on 06/09/2022).

[4] F. Ilievski, P. Szekely, and B. Zhang, "CSKG: The CommonSense Knowledge Graph," arXiv, Tech. Rep. arXiv:2012.11490, Mar. 2021, arXiv:2012.11490 [cs] type: article. DOI: 10.48550/arXiv.2012.11490. [Online]. Available: http://arxiv.org/abs/2012.11490 (visited on 06/08/2022).

[5] *Introducing the Knowledge Graph: Things, not strings*, en-us, May 2012. [Online]. Available: https://blog.google/products/search/introducing-knowledge-graph-things-not/ (visited on 06/08/2022).

[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*. USA: Prentice-Hall, Inc., 2009, ISBN: 978-0-13-187321-6.

[7] peterwestai2, *Symbolic Knowledge Distillation*, original-date: 2021-10-13T02:08:39Z, Jun. 2022. [Online]. Available: https://github.com/peterwestai2/symbolic-knowledge-distillation (visited on 06/09/2022).

[8] *Pretrained models — transformers 2.2.0 documentation*. [Online]. Available: https://huggingface.co/transformers/v2.2.0/pretrained_models.html (visited on 06/09/2022).

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," en, p. 24,

[10] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, *ZeRO: Memory Optimizations Toward Training Trillion Parameter Models*, Number: arXiv:1910.02054 arXiv:1910.02054 [cs, stat], May 2020. DOI: 10.48550/arXiv.1910.02054. [Online]. Available: http://arxiv.org/abs/1910.02054 (visited on 08/05/2022).

[11] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge," en, Nov. 2018. DOI: 10.48550/arXiv.1811.00937. [Online]. Available: https://arxiv.org/abs/1811.00937v2 (visited on 06/14/2022).

[12] A. Wang, Y. Pruksachatkun, N. Nangia, *et al.*, "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems," en, p. 30,

[13] P. West, C. Bhagavatula, J. Hessel, *et al.*, "Symbolic Knowledge Distillation: From General Language Models to Commonsense Models," en, Oct. 2021. DOI: 10.48550/arXiv.2110.07178. [Online]. Available: https://arxiv.org/abs/2110.07178v1 (visited on 06/08/2022).

Appendix A

Background Material

Material in an appendix.

We plot an equation in figure A.1.

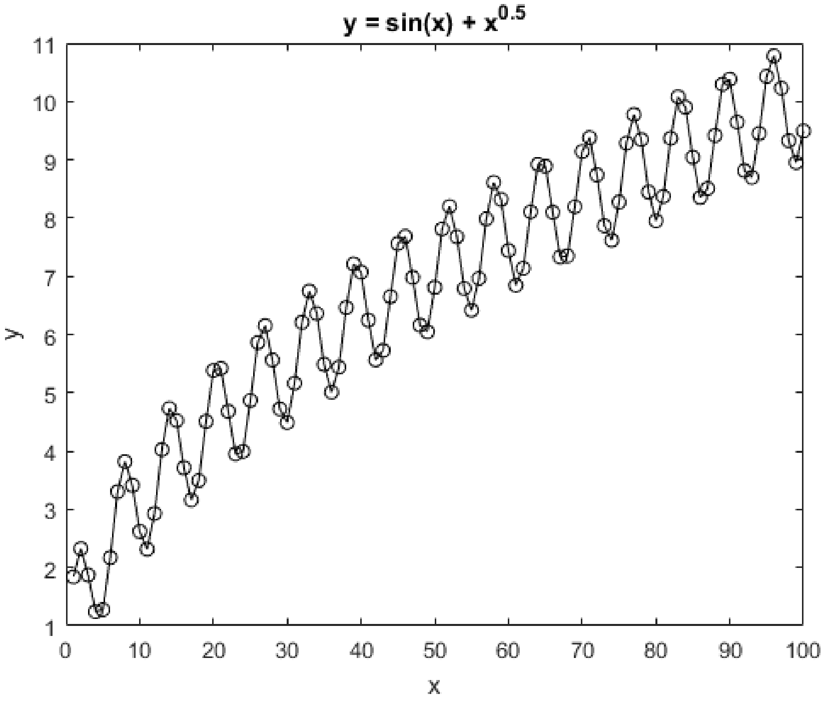

Figure A.1: A graph of $y = \sin(x) + \sqrt{x}$