Lu Sun, and many more.

A Notebook on Artificial Intelligence Updates

To my family, friends and communities members who have been dedicating to the presentation of this notebook, and to all students, researchers and faculty members who might find this notebook helpful.

Contents

Foreword vii

Preface ix

List of Figures xi

List of Tables xiii

I Transformer 1

1 Transformer Basics 3
- 1.1 It Seems That Attention is All You Need 3
- 1.2 Limitations to CNN and RNN 4
- 1.3 Attention Mechanism 4
- 1.4 Transformer 5
- 1.5 Latest Updates and Trends in Research of Transformer . . . 5

II Chatbot 7

2 OpenAI and Its Business Models 9

3 ChatGPT Basics 11

Bibliography 13

Foreword

If software or e-books can be made completely open-source, why not a notebook?

This brings me back to the summer of 2009 when I started my third year as a high school student in Harbin No. 3 High School. In around the end of August when the results of Gaokao (National College Entrance Examination of China, annually held in July) are released, people from photocopy shops would start selling notebooks photocopies that they claim to be from the top scorers of the exam. Much curious as I was about what these notebooks look like, never have I expected myself to actually learn anything from them, mainly for the following three reasons.

First of all, some (in fact many) of these notebooks were more difficult to understand than the textbooks. I guess we cannot blame the top scorers for being so smart that they sometimes make things extremely brief or overwhelmingly complicated.

Secondly, why would I want to adapt to notebooks of others when I had my own notebooks which in my opinion should be just as good as theirs.

And lastly, as a student in the top-tier high school myself, I knew that the top scorers of the coming year would probably be a schoolmate or a classmate. Why would I want to pay that much money to a complete stranger in a photocopy shop for my friend's notebook, rather than requesting a copy from him or her directly?

However, things had changed after my becoming an undergraduate student in 2010. There were so many modules and materials to learn in a university, and as an unfortunate result, students were often distracted from digging deeply into a module (For those who were still able to do so, you have my highest respect). The situation became even worse as I started pursuing my Ph.D. in 2014. As I had to focus on specific research areas entirely, I could hardly split much time on other irrelevant but still important and interesting contents.

This motivated me to start reading and taking notebooks for selected books and articles, just to force myself to spent time learning new subjects out of my comfort zone. I used to take hand-written notebooks. My very first notebook was on *Numerical Analysis*, an entrance level module for engineering background graduate students. Till today I still have on my hand dozens of these notebooks. Eventually, one day it suddenly came to me: why not digitalize them, and make them accessible online and open-source, and let everyone read and edit it?

—

As most of the open-source software, this notebook (and it applies to the other notebooks in this series as well) does not come with any "warranty" of any kind, meaning that there is no guarantee for the statement and knowledge in this notebook to be absolutely correct as it is not peer reviewed. **Do NOT cite this notebook in your academic research paper or book!** Of course, if you find anything helpful with your research, please trace back to the origin of the citation and double confirm it yourself, then on top of that determine whether or not to use it in your research.

This notebook is suitable as:

- a quick reference guide;
- a brief introduction for beginners of the module;
- a "cheat sheet" for students to prepare for the exam (Don't bring it to the exam unless it is allowed by your lecturer!) or for lecturers to prepare the teaching materials.

This notebook is NOT suitable as:

- a direct research reference;
- a replacement to the textbook;

because as explained the notebook is NOT peer reviewed and it is meant to be simple and easy to read. It is not necessary brief, but all the tedious explanation and derivation, if any, shall be "fold into appendix" and a reader can easily skip those things without any interruption to the reading experience.

—

Although this notebook is open-source, the reference materials of this notebook, including textbooks, journal papers, conference proceedings, etc., may not be open-source. Very likely many of these reference materials are licensed or copyrighted. Please legitimately access these materials and properly use them.

Some of the figures in this notebook is drawn using Excalidraw, a very interesting tool for machine to emulate hand-writing. The Excalidraw project can be found in GitHub, *excalidraw/excalidraw*.

Preface

Artificial intelligence (AI) was originally a research topic under the scope of control systems, and it is primarily used for system identification ("machine learning was used to be called system identification", as some professors say). Its artificial neuron network (ANN) structure is a promising approach for building highly nonlinear and data driven self-tuning functions.

Due to the limitation of computational resources in the past years, training ANN with massive size was hardly possible. With the advancement of computer and material science in the beginning of the 21st centry, in particular the development of GPU and TPU, nowadays we can manage networks with dozens of layers, each containing hundreds of neurons. This is referred as the deep learning (DL) network structure. With DL as the building block, we have invented convolutional neural networks (CNN) and recurrent neural networks (RNN), which are effective at finding trends in spacial and time-based data, respectively. The proposition of transformers, yet another attention-based deep learning structure, has taken natural language processing to the next level.

AI has grown so significantly in the past decades that it is now considered a separate area apart from control systems, and it is drawing more and more attentions than control systems. To put it in perspective, while a classic textbook on control systems may have 10,000+ citations records after years since its publication, a famous conference paper on modern AI can easily have 50,000+ citations after a few months.

This notebook does not focus on the introduction of basic AI mechanisms, as they are already covered in control systems related notebooks. The purpose of this notebook is mainly to keep up-to-date with the latest technology in AI, starting from transformer and its applications in chatbot, the most famous of which being ChatGPT developed by OpenAI.

Some part of this notebook is written by ChatGPT-4.

List of Figures

List of Tables

Part I

Transformer

1

Transformer Basics

CONTENTS

1.1 It Seems That Attention is All You Need 3
1.2 Limitations to CNN and RNN 4
1.3 Attention Mechanism 4
1.4 Transformer 4
1.5 Latest Updates and Trends in Research of Transformer 5

This chapter introduces the development of “transformer” in its main branch. Transformer is an AI framework initially proposed in 2017 in a paper “Attention is All You Need”, and it has magically revolutionized how we train AI for natural language processing and interpreting.

1.1 It Seems That Attention is All You Need

It started in 2017 when “transformer” is proposed in a paper “Attention is all you need”, by Vaswani, et al., a bunch of guys mainly from Google.

“Attention” was not a new concept previous to this paper. It was proposed a few years earlier than the transformer, and had been integrated into several CNN and RNN models as a supporting role to help boost their performances to capture long-range dependencies. Hence, it is quite surprising indeed, that when getting rid of CNN and RNN, and purely relying on attention mechanism, everything actually gets better: it is faster to train the network, and its performance gets improved.

This concludes that when getting rid of RNN and relying on attention mechanism alone, natural language processing actually gets better. Does this mean the RNN, which we had been relying on so much, is in fact the drawback to natural language processing? The following is the answer of ChatGPT-4 of this question.

Q: By getting rid of RNN and using attention mechanism alone, natural language processing performance gets better. Is this correct? Does this mean that RNN is a drawback to natural language processing?

A: Yes, by getting rid of RNNs and using the attention mechanism alone, the Transformer model demonstrated improved performance in natural language processing tasks. However, it's important to note that RNNs are not inherently a drawback to NLP; they were successfully used for many tasks before the advent of Transformers.

RNNs, particularly LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) variants, were designed to handle sequences and capture dependencies in the data. They have been effective in many NLP tasks, such as language modeling, machine translation, and sentiment analysis. However, RNNs do have certain limitations, such as difficulty in capturing long-range dependencies and sequential processing, which can lead to slower training and inference times.

The attention mechanism, as employed in the Transformer model, addresses some of these limitations. It enables parallel processing of input sequences, leading to faster training and inference, and effectively captures long-range dependencies without the vanishing gradient problem that can affect RNNs.

While Transformers have shown superior performance in many NLP tasks, RNNs still have their place in certain scenarios, especially when working with resource-constrained environments or smaller datasets. They can be more lightweight and computationally efficient compared to large-scale Transformer models.

In summary, RNNs are not a drawback to NLP per se, but their limitations have been mitigated by the introduction of the attention mechanism in Transformer models, which has led to improved performance in a wide range of NLP tasks.

(April 14, 2023, ChatGPT-4)

1.2 Limitations to CNN and RNN

“nobreak

1.3 Attention Mechanism

“nobreak

1.4 Transformer

“nobreak

1.5 Latest Updates and Trends in Research of Transformer

Part II

Chatbot

2

OpenAI and Its Business Models

CONTENTS

3

ChatGPT Basics

CONTENTS

Bibliography