山本レポート：システム安全性向上―世界の最前線（8）

大規模教師なし言語モデルの脅威

国立大学法人 名古屋大学
大学院 情報学研究科
教授　山本 修一郎
博士（工学）

今回は、OpenAI の大規模教師なし言語モデル GPT-2 の実験成果について紹介する[1-3]。この言語モデルでは、タスクに特化した学習の必要がなく、首尾一貫した複数の段落からなる一連の文章を生成でき、初歩的な文章理解、機械翻訳、質問応答、要約ができる。

GPT-2

GPT-2 は、800 万 Web ページのインターネット文章（40GB）から次の文章を予測するように訓練されるだけである。GPT-2 には、15 億個のパラメータがあるという。4B が 1 語だとすると、40GB は 100 億語になるから、15 語で 1 文が構成されると仮定すると、約 6 億 7 千万文の訓練データがあるともいえる。

GPT-2 のモデルは、与えられた文章に含まれる単語から次の単語を予測するという単純なものである。GPT-2 では、先代の GPT からパラメータと訓練データセットをそれぞれ 10 倍以上にしている。

GPT-2 は、ドメイン固有の訓練データセットを使用していないのに、特定のドメインについて訓練された他の言語モデルよりも優れているというので驚く。逆に言えば、分野独立な文章構成法があり、分野依存な単語関係があることを示唆していると考えられる。

GPT-2 は、分野に依存しない十分なデータセットにある生の文章を解析するだけである。

人間が文章を入力すると、後続する文章を自動的に生成してくれる。生成された文章はかなり説得性の高いものになっている。

具体例

GPT-2 が生成したということで、次の例を紹介している。

【GPT-2への入力例】

アンデス山脈のこれまで人が入らなかった谷で、科学者がユニコーンの群れを発見した。驚くべきことに、ユニコーンは完璧な英語を話した。

【GPT-2への出力例の冒頭部分】

科学者は彼らの独特の角、Ovid のユニコーンにちなんで、その集団を命名した。これらの四角い、銀白色のユニコーンは、これまで科学に知られていなかった。

今、約2世紀後に、この奇妙な現象を引き起こしたものの謎はついに解決された。

ラパス大学の進化論的生物学者である Jorge Perez 博士と数人の仲間は、他の動物や人間がいない小さな谷を見つけたときにアンデス山脈を探索していました。ペレスは、谷が自然の噴水であるように見え、岩と銀の雪の2つの山に囲まれていることに気づいた。

このあとに、さらに6つの段落が生成されたと紹介されている。実に、もっともらしい文章だ。この文章を見せられて、一般人がこの文章の嘘を見抜くことは不可能だ。普通の人は信じてしまうだろう。

このように、GPT-2 を使えば、スパムメール、フェイクニュースや不正論文を簡単かつ大量に自動生成できてしまう可能性がある。これはソーシャルメディアだけでなく、学術界にとっても大きな弊害を生むことになる。

現在の学会の論文査読システムは、査読者が論文を審査している。もし機械的に論文が作成されてしまうと、人間による査読システムでは追い付かないし、不正を簡単には見破ることが難しくなるだろう。

限界

とはいえ、現在の GPT-2 には、次のような限界があることも紹介している。

世界のモデル化の失敗

たとえば、水中で起こる火災など不自然なトピックの切り替え。

世界をモデル化するのは、自然言語処理分野では、古い問題である。人間にとって常識的な事柄をモデル化するのは GPT-2 の人工知能にとって困難な研究対象である。現在の機械学習の限界は、常識のモデル化であることを GPT-2 も示しているわけだ。結局、常識のモデル化は AI を開

発する人間の仕事ということだ。

トピックの切り替えについては、現在の GPT-2 が使用している 800 万 Web ページでは不足しているということだ。また、GPT- 2 が教師なし機械学習による言語モデルの限界でもある。トピックの切り替えについて GPT- 2 が教師データを使用すれば改善できそうである。ただし、やはり、膨大な規模の教師データが必要になるのではないかと思われる。この理由は、トピックの切り替えは文章の文脈に依存するからだ。800 万 Web ページのデータセットでも自然なトピックの切り替えが不完全なのだから。

GPT-2 が適切な文章を生成する確率は現在のところ、50%だという。

文章の内容が高度に技術的または難解な場合は、モデルが生成する文章の品質が低下する可能性があるようだ。この対策として、たとえば、Amazon レビューの文章をデータセットとして GPT-2 を微調整することにより、星の評価やカテゴリなどの条件でレビュー文章を生成できるとしている。しかし、この方法は、GPT-2 の適用分野を Amazon レビューに限定することだから、GPT-2 が目指した汎用的な文章を生成する方法を修正することでもある。つまり、汎用的なデータセットだけでは、結局のところ特定分野における文章がうまく生成できる可能性は半分ということだ。とはいえ、2 回の自動生成で 1 回はうまくいくともいえるから、大したものである。自分で考えるよりはよほど効率的かもしれない。

何か問題を起こした時に、もっともらしい言い訳を書くのには、役に立ちそうだ。

GPT-2 が示唆すること

GPT-2 のような大規模教師なし言語システムは次のような分野に適用できる。

- **文章作成支援**
- **対話エージェント**
- **翻訳支援**
- **音声認識**

GPT-2 のような大規模教師なし言語システムは次のような不適切なシステムの構築に適用できる。

- **フェイクニュースの作成**
- **なりすまし文章の作成**
- **偽造文章の作成**
- **スパム／フィッシング文章の作成**

敵意のある行為者が大規模教師なし言語技術を用いて真贋性の確認が困難な文章を大量に生成する機会が増える可能性が高くなっている。したがって、このような不正文章を除去するための技術開発と非技術的対策の立案が急務である。いずれにしろ、どのような技術にも明暗がある。より高度な技術はより巧妙に悪意される可能性があるということである。機械学習もインターネットも同じである。このような悪用を防ぐには人間が倫理的な規則を用いて制御することが不可欠である。

GPT-2 の公開方針

前述したことから、GPT-2 が驚異的な成果を達成したことで、逆に GPT-2 がもたらす社会的な影響が脅威になったと開発者が感じたのだろう。GPT- 2 のデータセット、トレーニングコード、モデルの重みについては、公開していない。

OpenAI 憲　章（https://blog.openai.com/openai-charter/）では、「安全性、ポリシー、および規格に関する研究を共有することの重要性を増しながら、将来的に安全性とセキュリティの懸念が従来型の出版を減らすことが予想される」と書かれている。今後どのような出版（文章の作成）のあり方が望ましいかについての議論が必要だとしている。

まとめ

- 大規模教師なし言語モデルは、ドメイン固有のデータセットで訓練されたモデルより優れている。
- 大規模教師なし言語モデルによって、不正文章を大量に生成できることから、社会的な脅威が増大する。
- 責任を持って出版する（文章を作成する）ことをどのようにして保証するかが問われている。著者が自分で考えたこと、つまり AI で自動生成していないことを証明するような認証マークが必要になるかもしれない。

【参考】
[1]Better Language Models and Their Implications, FEBRUARY 14, 2019 https://blog.openai.com/better-language-models/
[2]ITmedia NEWS, OpenAI, まことしやかなフェイクニュースも簡単生成の言語モデル「GPT-2」の限定版をオープンソース化, 2019年02月15日 10時43分 公開, https://www.itmedia.co.jp/news/articles/1902/15/news075.html
[3]CNET Japan, マスク氏が支援するOpenAI、大規模な教師なし言語モデル「GPT-2」の情報を公開, https://japan.cnet.com/article/35132788/

＜システム安全性のことなら下記へ＞
yamamotosui@icts.nagoya-u.ac.jp