Model Card for CTRL - A Conditional Transformer Language Model for Controllable Generation

The CTRL Language Model analyzed in this card generates text conditioned on control codes that specify domain, style, topics, dates, entities, relationships between entities, plot points, and task-related behavior.

On this model card, you can learn more about how this model was trained, its capabilities, its intended use, and its limitations.

Model Details

Organization
Salesforce Research

Model date	***Model type***	***Input***
September 10, 2019	Language model	Text

Information about parameters
1.63 billion-parameter transformer language model

Output
The model can generate text conditioned on control codes that specify domain, style, topics, dates, entities, relationships between entities, plot points, and task-related behavior

Read the full paper here:
https://arxiv.org/abs/1909.05858

Access the public code here:
https://github.com/salesfore/ctrl

Citation details
Title: CTRL - A Conditional Transformer Language Model for Controllable Generation
Authors: Keskar, Nitish Shirish and McCann, Bryan and Varshney, Lav and Xiong, Caiming and Socher, Richard
Journal: arXiv preprint arXiv:1909.05858
Year: 2019

License: BSD 3-Clause (https://github.com/salesforce/ctrl/blob/master/LICENSE.txt)

Metrics

Model performance measures
Performance evaluated on qualitative judgments by humans as to whether the control codes lead to text generated in the desired domain

Training Data

This model is trained on 140 GB of text drawn from a variety of domains: Wikipedia (English, German, Spanish, and French), Project Gutenberg, submissions from 45 subreddits, OpenWebText, a large collection of news data, Amazon Reviews, Europarl and UN data from WMT (En-De, En-Es, En-Fr), question-answer pairs (no context documents) from ELI5, and the MRQA shared task, which includes Stanford Question Answering Dataset, NewsQA, TriviaQA, SearchQA, HotpotQA, and Natural Questions. See the paper for the full list of training data

Intended Use

Primary intended use

1. Generating artificial text in collaboration with a human, including but not limited to:
   - Creative writing
   - Automating repetitive writing tasks
   - Formatting specific text types
   - Creating contextualized marketing materials
2. Improvement of other NLP applications through fine-tuning (on another task or other data, e.g. fine-tuning CTRL to learn new kinds of language like product descriptions)
3. Enhancement in the field of natural language understanding to push towards a better understanding of artificial text generation, including how to detect it and work toward control, understanding, and potentially combating potentially negative consequences of such models

Primary intended users
General audiences
NLP researchers

Out-of-scope use cases

- CTRL should not be used for generating artificial text without collaboration with a human.
- It should not be used to make normative or prescriptive claims.
- This software should not be used to promote or profit from:
  - violence, hate, and division;
  - environmental destruction;
  - abuse of human rights; or
  - the destruction of people's physical and mental health.

Ethical Considerations

We recognize the potential for misuse or abuse, including use by bad actors who could manipulate the system to act maliciously and generate text to influence decision-making in political, economic, and social settings. False attribution could also harm individuals, organizations, or other entities. To address these concerns, the model was evaluated internally as well as externally by third parties, including the Partnership on AI, prior to release.

To mitigate potential misuse to the extent possible, we stripped out all detectable training data from undesirable sources. We then red-teamed the model and found that negative utterances were often placed in contexts that made them identifiable as such. For example, when using the 'News' control code, hate speech could be embedded as part of an apology (e.g. "the politician apologized for saying [insert hateful statement]"), implying that this type of speech was negative. By pre-selecting the available control codes (omitting, for example, Instagram and Twitter from the available domains), we are able to limit the potential for misuse.

In releasing our model, we hope to put it into the hands of researchers and prosocial actors so that they can work to control, understand, and potentially combat the negative consequences of such models. We hope that research into detecting fake news and model-generated content of all kinds will be pushed forward by CTRL. It is our belief that these models should become a common tool so researchers can design methods to guard against malicious use and so the public becomes familiar with their existence and patterns of behavior.

Caveats and Recommendations

- A recommendation to monitor and detect use will be implemented through the development of a model that will identify CTRL-generated text.
- A second recommendation to further screen the input into and output from the model will be implemented through the addition of a check in the CTRL interface to prohibit the insertion into the model of certain negative inputs, which will help control the output that can be generated.
- The model is trained on a limited number of languages: primarily English and some German, Spanish, French. A recommendation for a future area of research is to train the model on more languages.