Model Card: Detector for CTRL from Salesforce Research

Basic Information

Neural language models like CTRL and GPT-3 are now capable of generating text that is difficult to distinguish from human-authored content. These models can be used to facilitate creative writing, automate repetitive writing tasks, or create contextualized marketing materials. But they can also undermine user trust by generating and propagating fake news, setting political agendas, and even influencing elections. It can be difficult for humans to detect this machine-generated text. For that reason, Salesforce has released a companion CTRL detector, which reports the probability that text was generated by CTRL (or not). The research examined the capability and limits of automatic detection of machine-generated text.

Model Details

Organization	Salesforce, Inc.
Model date	October 13, 2020
Model version	1.0
Model type	Classifier
Training data, citations	RoBERTa CTRL GPT-2 WebText RealNews Trump Speech
License	BSD3
Github	www.github.com/MetaMind/CTRL-detector
Send questions or comments to:	ai_ethics@salesforce.com

Intended Use

Primary intended use

1. Assist NLP researchers in improving detector models for this or other machine-generated models.
2. Provide social media platform moderators and consumers the probability that a selection of text is generated by CTRL, with an aim to detect machine-generated:
    - fake news;
    - texts that manipulate political opinions; and/or,
    - hate speech.

Primary intended users

- NLP researchers (can be beneficial to further develop detector models)
- Social media platform moderators
- Civil society actors with an interest in identifying and countering mis- and disinformation

Out-of-scope use cases

- The CTRL detector should not be used for the adversarial training of fake news generators.
- 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.
- The CTRL detector should not be used to automate a production-level system - this is currently in a state of applied research intended for the audiences noted above.

Metrics

Model performance measures

The detection rate is defined as the number of correct predictions (machine-generated vs. human-generated) over the total number of samples. The decision threshold is 0.5 predicted probability.

Three types of evaluations were conducted:

1. Detecting machine-generated text of the same type on which CTRL was trained (most basic scenario, deploying the CTRL detector on the same distribution of human and machine-generated text from the same WebText domain).
2. Detecting machine-generated text fine-tuned to something fairly generic (medium-case, deploying the CTRL detector to identify text trained on RealNews, with a slightly different distribution than WebText).
3. Detecting machine-generated text fine-tuned to a very specific domain (most difficult case, using Trump tweets).

Accuracy varied depending on how similar the text to be detected was to the text on which the CTRL detector was trained.

Training Data

The training dataset consists of 250k documents from the English language Webtext test set and 250k documents generated from CTRL and GPT-2 Language models. For the language models, we generated samples using Top-K 40 truncation and 3 words of priming. The priming words were extracted from Webtext distinct from the 250k samples used in the training dataset. The texts were trimmed to length of 256 tokens using the BPE (Byte Pair Encoding) tokenizer.

Quantitative Analyses

To begin, we fine-tuned a RoBERTa model on the task of labeling a sentence as human- or machine-generated. From the results reported in Table 1, we found that detection rate drops as the length of a test sentence drops, while it increases if the training data for the detector's length is similar to that of the test data.

Table 1: Automatic detection rates on CTRL-generated test set:

Method	**Test Length - 32 tokens**	**Test Length - 16 tokens**	**Test Length - 8 tokens**
Train on Length 256	91.97	72.26	57.82
Train on Test Length	95.77	91.09	68.31

Table 2: Automatic detection rates on the CTRL (Greedy) - generated data set:

Method	**Detection Rate**
Train on CTRL (Greedy)	99.86
Train on GPT-2 (Top-40)	91.49 (-8.37)
Train on CTRL and GPT-2	99.40

Table 3: Automatic detection rates on the GPT-2 (Top-40) - generated data set:

Method	**Detection Rate**
Train on GPT-2 (Top-40 Sampling)	96.11
Train on CTRL (Greedy Sampling)	50.11 (-46.11)
Train on CTRL (Top-40 Sampling)	56.89 (-39.22)
Train on CTRL and GPT-2	99.12 (+3.01)

As demonstrated in Tables 2 and 3, the CTRL detector generalizes relatively well to GPT-2 on CTRL-generated texts, whereas the detector generalizes poorly to GPT-2 generated texts. This may be due to the fact that GPT-2 is trained only on Webtext data (similar to test data), while CTRL is trained on multiple text sources. Therefore, it is possible that the CTRL detector might overfit the distributional difference between training data of CTRL and Webtext. Also, we observe that the CTRL detector trained on text generated with the same sampling method as the test set performs better.

Finally, we evaluated the CTRL detector's accuracy and generalizability across fine-tuned models. To examine this performance, we conducted experiments on two fine-tuned datasets: Real News (general domain, 10k) and Trump Speech (specific domain, 10k Tweets)

Real News is a dataset extracted using Common Crawl from the following websites: theguardian.com, reuters.com, nytimes.com, theatlantic.com, usatoday.com, huffingtonpost.com, and nbcnews.com.

Table 4: Automatic detection rates on the CTRL (fine-tuned)-generated data:

Method	Fine-tuned on Trump Speech	Fine-tuned on Real News	No fine-tuning
Train on CTRL	80.71	94.97	99.86
Train on GPT-2	90.41	92.25	91.49
Train on CTRL and GPT-2	95.93	99.26	99.40

From this, we observe that detection rate drops as the detector is applied to texts generated by a fine-tuned generator. An important observation is that training on multiple generators helps detection on sentences generated by fine-tuned models.

Ethical Considerations

We hope this research may assist in detecting fake news on social platforms in the hands of good actors. However, we recognize the potential for the CTRL detector's misuse or abuse, including use by bad actors who might use it to adversarially train a fake news generator, with the attendant potential to influence decision-making in political, economic, and social settings. While there is no specific mitigant to this risk, our hope is that the positive contribution of this research, to the ability to automatically detect machine-generated speech far outweighs the risks.

In releasing the CTRL detector, 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 text generation models. We also hope that its release helps to push forward further research into detecting fake news and model-generated content of all kinds.

Caveats and Recommendations

- Training data was truncated to between 8 and 256 tokens - further research on tokens both below and above this threshold may be merited.
- This model is not yet ready for production. Its development was intended to push forward the science and research into human-generated language and the ability to distinguish machine-generated text. Further research is recommended and needed. Similarly, it should not be assumed that the CTRL detector's prediction is sufficient to remove or flag a piece of content in a production-level system. Human review is both merited and useful.
- The detector model is trained only on an English language corpus. We recommend future research into training the model on non-English languages.
- The CTRL detector has only been trained on CTRL and GPT-2 and may not be generalizable to unseen generators (one to which we do not have access, with, for example, the same architecture but trained on a different data source). Further research is recommended to create a detector that may be more generalizable, to ascertain how it would perform on a generator model that was not itself used to generate the training data.