Derek Lee Deep Learning Fall 2020

Professor Curro Quiz #9

1. Identify something you find "unsavory" about these kinds of models; explain and argue.

The fact that these models are susceptible to biases in the training data, and that the training data was taken from the internet, is unsavory. For example, “Language Models are Few-Shot Learners” mentions several experiments conducted to examine the biases present in the model. The authors show that the model has gender, racial, and religious biases, among others. This is due to the training data used for the model. The training data consists of human writing, which contains the biases of the people who wrote it. Any existing human biases in the data will also be present in the model. The data was taken from the internet, which has “internet-scale bias”. In addition, the pre-trained versions of these models are widely used as the base for other, more specialized models. The amount of training data and training time required to train these models make training from scratch impractical. Thus, specialized models are created through few-shot learning on the pre-trained models. This means that those specialized models will carry over the biases present in the pre-trained models.

It is necessary to actively intervene to counter these biases. However, it is difficult to do so. The method used by the authors to demonstrate bias required testing the model on specific phrases, which were carefully crafted to discern any potential bias. Scaling that method will be difficult. In addition, characterizing the bias is difficult as well. For example, when the authors explicitly prompted the model to discuss race, they acknowledge that “the resulting sentiment can reflect socio-historical factors… text relating to a discussion of slavery will frequently have a negative sentiment, which may lead to a demographic being associated with a negative sentiment” under the testing methodology used. The authors note that “mitigation work should not be approached purely with a metric driven object to ‘remove’ bias as this has been shown to have blind spots but in a holistic manner.”

Something else that is unsavory is that “Language Models are Few-Shot Learners” suggests that it may be possible that few-shot learning does not actually learn new tasks “from scratch”, but instead “recognizes and identifies tasks that it learned during training.” These types of models are very “expensive and inconvenient to perform inference on”. They contain a “very wide range of skills, most of which are not needed for a specific task”. It is possible that these models can be reduced in size, without a significant reduction in quality. In fact, the suggestion that the models contain a “very wide range of skills” supports the idea that few-shot learning only selects tasks that were already learned during training, rather than learning new ones.