Comparison and Contracts Between GPT and GPT-2 Transformer Language Models

OpenAI has introduced generative pre-trained (GPT) models, which do the language processing in unsupervised manner. Until their introduction, the language modelling tasks e.g. translation, summarization, question-answering etc. used to be done by doing the supervised training the language models to do the specific tasks.

GPT came in mid 2018 and in early 2019 came GPT-2 model. GPT-3 has also been announced in 2020 but that is out of the scope of our discussion.

The early GPT model showed that if the large training corpus has been used for unsupervised training for a language model, it can be trained a little in the supervised way for language specific tasks e.g. textual entailment, semantic similarity, reading comprehension, sentiment analysis, linguistic acceptability, multi-task benchmark, question-answering etc. The GPT-2 was trained in unsupervised way on 10 times of data volume of GPT and had 10 times as many parameters as in GPT. GPT-2 demonstrated that with such a huge pre-trained dataset, no language task specific training is needed and yet the model will either achieve task specific state-of-the-art (SOTA) performance or come very close to it. Therefore, the GPT transformer-based language model will be generic solution to every possible language use-case.

In the next paragraphs, I'll discuss GPT in more detail, its drawbacks and how GPT-2 supersedes, along with the gist of the controversy the non-publication of GPT generated in the research community.

GPT extended the concept of semi-supervised sequence learning, which showed how to improve the document classification performance by using unsupervised pre-training and supervised fine-tuning. This transformer-based model can then be used on a broad range of language-based use-cases and thus will work as a generic language model. The model does require little bit of supervised training but very little or no hyperparameter tuning is needed for language specific tasks.

The makers of GPT didn't had full understanding as to how generative pre-training helps. They also wanted to make improvement in the fine-tuning part of the GPT model by adaption and transfer techniques.

It was noticed during GPT research that for many of the tasks e.g. question-answering in the multiple-choice questions, the performance increases when the underlying model improves without any need of the training of the model of specific use-case. Therefore, the natural extension of the GPT was to keep training it on more data and keep improving the robustness of the transformer model. GPT-2 was the direct extension of this process.

GPT-2 is trained with the objective of predicting the next word, given the previous words in the text. GPT-2 is available in three versions by OpenAI, depending upon the training dataset size and the model parameters. The largest version is trained on 40 GB of text data and has 1.5 billion hyperparameters.

GPT-2 has broad range of capabilities including generation of the lengthy continuation of the text of the unprecedented quality. It surpasses other languages models, which are trained on specific language domains.

The OpenAI envisaged that GPT-2 would've many societal advantages e.g. language translation, writing assistance, speech recognition etc. GPT-2 has indeed had profound impact in democratization of SOTA pre-trained language models for variety of use-cases.

One very interesting feature of GPT models is that they are trained on the publicly available text corpus and then tested on the same datasets on which the previous SOTA language models have based their SOTA accuracy, therefore, making the comparison between GPT and rest of the models straight-forward. The GPT itself surpassed many SOTA models of its time and gave competitive results in rest of the use-cases. GPT-2 gave higher accuracy on pre-measured stellar results from GPT, due to its sheer size of the pre-trained data.

The fact that GPT requires very little last layer supervised training for task specific use-cases and GPT-2 doesn't require any supervised training means there is a lot of time and resource saving, which would've gone in supervised training. Data collection, cleaning and curation takes huge amount of time and unsupervised pre-training of the model helps to get rid of this effort.

GPT and GPT-2 both enable the transfer learning in the sense that these models can be used for any of the language specific use cases with very little or no training, therefore, the initial training done by OpenAI is being transferred to anyone using the model later. Even though pre-training of the GPT and GPT-2 models takes huge amount of resources, the transfer learning compensates that and makes the pre-trained model available to everyone.

GPT-2 has potential to be misused for malicious purposes e.g. online impersonation, generating misleading news, automate fake content production, creating biased and abusive content etc., therefore, OpenAI released only the very compact version of GPT-2 without any dataset and code. This created a major controversy in AI community, which thrives on the openness and research collaboration. The opponents of OpenAI's stand sited it as indirect monetary advantage tactic (e.g. so that more papers can be produced) whereas the proponents supported the ethical stand taken by OpenAI until the regulatory measures by the governments are put in place. Since then the intermediate version of the GPT-2 model has been published to the general community but the controversy still rages on.

The transformer paper 'Attention is All You Need' has fundamentally changed the language modelling for NLP (natural language processing). GPT and GPT-2 which are based on transformer mechanism have proved that unsupervised pre-trained language models work as generic solution to all language-based use-cases and no supervised training is needed. This has led to explosive growth in NLP, not only in research but in industry as well. Since then GPT-3 has been published, which is biggest of any language model till date and simply outperforms every SOTA model in most of the language-based use-cases.

References:

1. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
2. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
3. https://openai.com/blog/better-language-models/#task6
4. https://openai.com/blog/language-unsupervised/
5. https://www.youtube.com/watch?v=SY5PvZrJhLE
6. https://www.youtube.com/watch?v=u1_qMdb0kYU