Resumo – Language Models are Few-Shot Learners

O artigo apresenta Generative Pretrained Transformer (GPT-3), um modelo de linguagem autorregressivo de 175 bilhões de parâmetros, dez vezes maior que qualquer outro modelo não esparso anterior. Trabalhos anteriores demonstraram ganhos substanciais em tarefas de NLP por meio do pré-treinamento agnóstico em um grande corpus de texto seguido de ajuste fino em tarefas específicas, porém, que requer grande massa de dados rotulados, e tem potencial de não generalizar fora da distribuição. Por outro lado, os humanos geralmente podem realizar uma nova tarefa de linguagem a partir de apenas alguns exemplos ou de instruções simples. A proposta do artigo é mostrar que escalando enormemente modelos de linguagem treinados agnosticamente consegue-se melhorar o desempenho few-shot, algumas vezes alcançando competitividade com melhores modelos anteriores com ajuste fino.

Durante o pré-treinamento não supervisionado, um modelo de linguagem desenvolve um amplo conjunto de habilidades e a capacidade de reconhecimento de padrões, e usa essas habilidades no momento de inferência para se adaptar rapidamente à tarefa desejada, o que é chamado de aprendizagem no contexto. Curvas de aprendizagem no contexto mais íngremes demonstram que modelos bem maiores possuem habilidades superiores de aprender tarefas a partir do contexto.

A avaliação do GPT-3 foi realizada sobre 24 conjuntos de dados para NLP, e várias tarefas projetadas para testar a rápida adaptação a tarefas que não possam ser inferidas diretamente do dado de treinamento. Para cada tarefa, avalia-se o GPT-3 sob 3 condições: aprendizado few-shot; aprendizado one-shot; e aprendizado zero-shot. No experimento, não se avalia o desempenho do GPT-3 em ajuste fino tradicional. São testadas 8 configurações de modelos com diferentes tamanhos do GPT-3 nas condições descritas.

Em tarefas de NLP em geral, o GPT-3 alcançou resultados excelentes nas condições zero-shot e one-shot, sendo algumas vezes competitivo ou eventualmente até superando o estado da arte em condições few-shot. GPT-3 também mostra proficiência one-shot ou few-shot em tarefas projetadas para rápida adaptação ou raciocínio em tempo real, como desembaralhar palavras, efetuar aritmética, e empregar palavras novas após tê-las visto uma única vez.

Ao mesmo tempo, em algumas tarefas, há problemas de desempenho few-shot mesmo em modelos na escala do GPT-3, como inferência no conjunto de dados ANLI, ou compreensão de leitura nos conjuntos de dados RACE ou QuAC. O GPT-3 apresenta desempenho pior em tarefas que se beneficiam empiricamente da bidirecionalidade. Uma possibilidade é repetir em modelo bidirecional a escala do GPT-3 com meta-aprendizagem. Também foram identificados vieses sociais no comportamento dos modelos, podendo gerar conteúdo estereotipado ou preconceituoso, quanto a sexo, raça e religião, por exemplo.

O artigo apresentou um modelo de linguagem de pré-treino massivo (175 bilhões de parâmetros) que apresenta que mostra um forte desempenho em muitas tarefas de processamento linguagem natural e benchmarks nas condições zero-shot, one-shot e few-shot, em alguns casos quase igualando o desempenho de sistemas finamente ajustados de última geração, além de gerar amostras de alta qualidade e forte desempenho qualitativo em tarefas definidas em tempo real. Apresentou possíveis tendências de escalamento do desempenho sem realizar ajustes finos. Também discutiu os impactos sociais dessa classe de modelo. Apesar de suas limitações e fraquezas, os resultados sugerem que modelos de linguagem massivos podem ser um ingrediente importante no desenvolvimento de sistemas de linguagem gerais adaptáveis.