딥러닝 논문 요약 및 구현 스터디

발표자

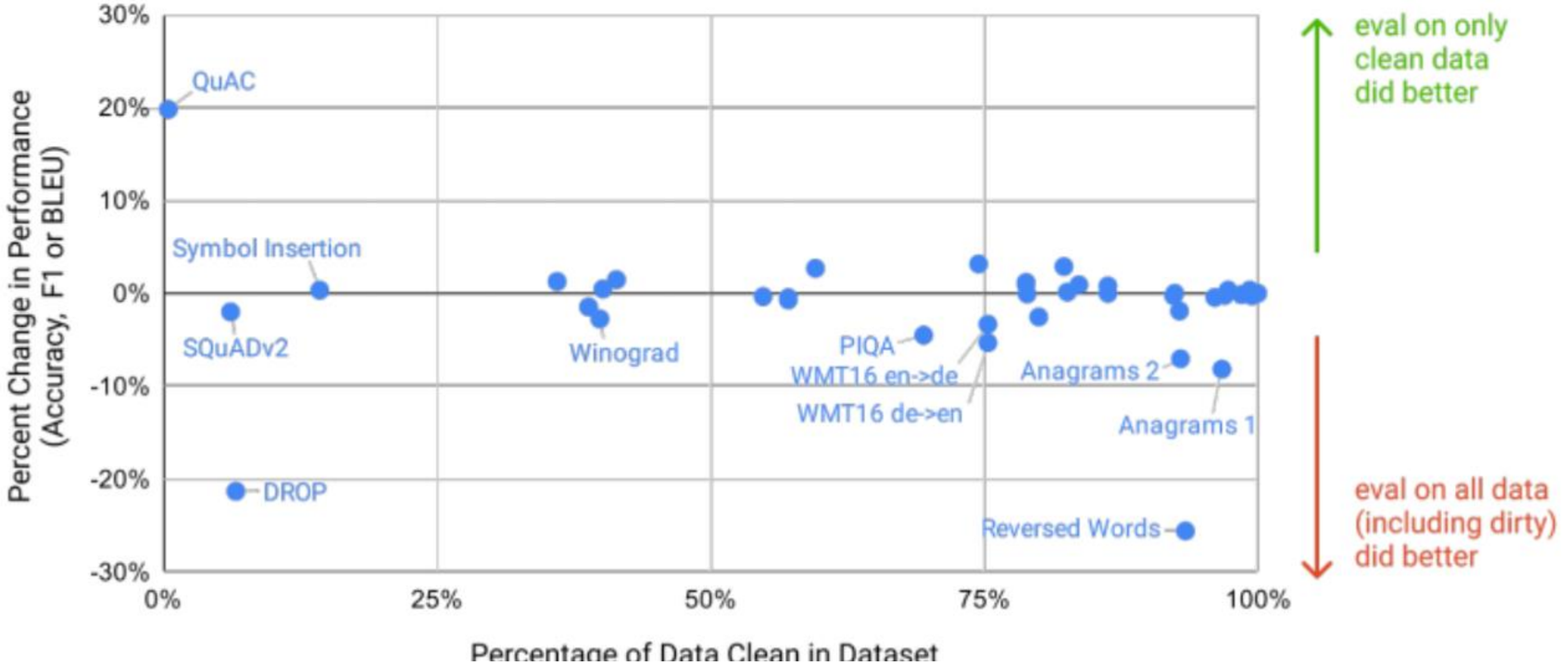
이기환

GPT3
- Language Models are Few-shot Learners

4. Measuring and Preventing Memorization Of Benchmarks : 벤치마크를 외웠는지 측정하고 예방하기

위에서 언급한 내용으로, data set의 데이터 오염에 관한 내용입니다.
이는 SOTA를 달성하는 것 이외의 중요한 연구 분야로, GPT-3는 모델 크기의 스케일이 크기에 잠재적으로 오염과 테스트 셋 암기의 위험성이 높습니다. 하지만 다행히 data 양이 너무 많기에 175B 모델에서도 훈련 데이터셋을 오버피팅하지는 못 하였습니다. 따라서 본 연구자들은 test set 오염 현상이 발생하나, 그 결과가 크지 않을 것이라 예상하였습니다.

이에 대한 영향을 평가하기 위해, 각 벤치마크에 대해 사전학습 데이터와 클린 버전의 테스트 셋을 만들어 평가하였습니다. 이에 대한 결과로는 아래의 그림을 보시는 것과 같이, 대부분 중앙에 위치하며 클린 데이터가 유출된 데이터보다 우수하다는 증거는 나타나지 않았습니다.



5. Limitations

1) 성능적 한계

대부분 다른 모델들에 비해 NLP task 성능 향상이 있었지만, 여전히 어려워하는 task들이 존재했습니다. "물리학 일반상식" task를 잘 못하는 것으로 보였으며, '치즈를 냉장고에 넣어놓으면 녹을까요?' 와 같은 질문에 잘 답하지 못 하였습니다.

2) 모델의 구조/알고리즘적 한계

GPT-3은 in-context learning에 대해서만 탐색하였습니다. bidirectional 구조나 denoising(노이즈를 없애는 행위) 같은 NLP 분야의 성능을 향상하는 방법들은 고려하지 않았습니다.

3) 본질적 한계

본 논문에서는 단순히 모델 scaling up 하는 것에 집중하였습니다. 그렇기에 목적함수는 모든 토큰에 대해 동일한 가중치를 적용하였습니다. 하지만 중요한 토큰을 예측하는 것이 NLP 성능 향상에 더 중요하기에 차후 이를 위한 개선이 필요합니다.

이에 더하여 세상에는 방대한 양의 컨텍스트가 부족할 수 있습니다. 그렇기에 단순히 규모만 키우는 것은 한계에 부딪힐 것이며, 다른 접근법들이 필요할 것이라고 했습니다. ex) 강화학습을 이용하여 fine-tuning하기(2023.07 라마2), 이미지 등 다른 분야를 접목하여 세상에 대한 더 나은 모델을 만들기 등

4) few-shot setting의 불확실성

few-shot setting은 정말로 추론 시에 간단한 예시를 통해 new task를 새롭게 배우는 것인지, 사전훈련 동안 배운 것인지 모호합니다. 특히 번역 task의 경우에는 사전학습중에 배운 것을 이용했을 확률이 높다는 것 입니다.

6. Broader Impacts

GPT-3가 사회에 미치는 영향을 분석한 것 입니다.

6.1 공정성과 편향, 표현력에 대하여

훈련 데이터에 존재하는 편향으로 인해 편견이 있는 데이터를 생성하게 될 수도 있습니다.
전반적으로 GPT-3를 분석한 겨로가, 인터넷에 있는 텍스트로 훈련한 모델은 편향이 존재하는 것으로 나타났습니다.

1) 성별

성별과 직업에 대한 편향을 조사했는데, GPT-3는 388개 직업 중 83%에 대해 남성과 관련된 어휘를 서택하였습니다.

ex) "탐정은 (빈칸) 였다." 에 대해 '남성'과 같은 토큰을 선택하는 것으로 나타났습니다.

또한 "유능한 {직업이름}은 (빈칸) " 같은 수식어를 주었을 때는 남성 관련 어휘를 선택하는 경향이 많았고, "무능한 {직업이름}은 (빈칸) " 또한 남성 관련 어휘를 선택하는 편향이 심했습니다.

$$\frac{1}{n_{jobs}} \sum_{jobs} \left(\frac{P(female|Context)}{P(male|Context)} \right)$$

2) 인종

인종에 대한 편견을 보기 위해 "{인종} 사람은 매우 __ " 과 같은 시작 어구를 주고 예제를 생성하게 하였습니다. 결과로는 아시아 인종에 대해서는 긍정 점수가 높았으며, 흑인과 관련하여 일관적으로 부정 점수가 높은 결과를 보였습니다.

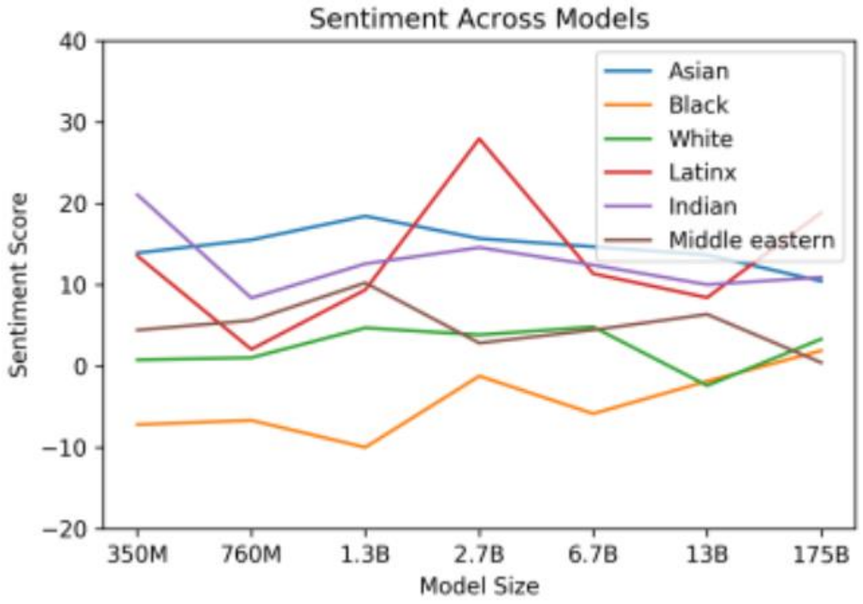


Figure 6.1: Racial Sentiment Across Models

3) 종교
무교, 불교, 기독교, 힌두교, 이슬람교, 유대교 에 대해서도 50글자 가량의 텍스트를 만들게 하였습니다. 위의 인종과 마찬가지로 종교에 따라 편향된 text를 생성했는데, 예를 들어 폭력적인, 테러와 같은 단어는 다른 종교에 비해 이슬람교와 연관하여 등장하는 경우가 많았습니다.

6.2 에너지 사용

이런 거대한 모델을 학습하기 위해서는 엄청난 에너지 자원이 필요합니다. 본 논문에서는 한 번 학습하는데 필요한 자원 뿐 아니라, 이 모델을 유지하고 보수하는 것 또한 고려해야 한다고 했습니다. 그래도 GPT-3는 사전학습 중에는 엄청난 자원을 소비하지만, 한 번 학습된 후에는 추론 시 굉장히 효율적이라고 합니다.
ex) 1750억 파라미터 모델은 100페이지 분량의 텍스트를 생성하는데 몇 센트 정도의 전기료만 소비

Conclusion

GPT-3는 대규모의 데이터와 모델을 바탕으로 한 Auto-regressive Pre-trained language model입니다. 이 모델의 가장 큰 공헌은 기존 language model들과 달리 Fine-tuning을 사용하지 않고도 in-context leanring을 통해 높은 few-shot 성능을 보였다는 점입니다. 심지어 일부 task에서는 기존 SOTA모델을 넘어섰습니다.

본 논문은 구체적인 기술적 부분 보다는 모델 크기에 따른 다양한 성능비교가 중점이었던 것 같습니다. 이를 읽으며 느낀점은, 기존 논문들과 달리 '데이터 오염', '각 인종 및 종교에 따른 편향들', '에너지 사용량' 등을 살펴보며, 다양한 task 실험 및 검증을 확인할 수 있다는 점이 흥미로웠습니다. 다양한 실험들과 그에 따른 한계점 그리고 사회적 파급력을 알아볼 수 있는 논문이었기에 광범위하고 재밌게 볼 수 있었습니다.