Attention is not not explanation

review

Tobig's 19-20기 자연어 심화세션 2조

20기 경민수

Main Points

Attention is not not explanation (2020, EMNLP)
- Sarah Wiegreffe & Yuval Pinter (Georgia Institute of Technology)

- Explanation can mean different things, ans J&W(attention is not explanation) are not clear on which interpretation they wish to disassociate from attention models
  - Explanation의 다양한 정의와 J&W의 불명확한 용어 정의

- Their(J&W) first empirical study, a correlation-based analysis of attention scores, is neither sufficient to advance the claim nor convincing in its results
  - J&W의 상관관계 기반 분석 연구의 설득 가능성 하락

- Their second study, an attention-distribution manipulation experiment, is orthogonal to the claim, and designed with such a high degree of freedom, that is results have little to no meaning.
  - 매우 높은 자유도로 설계되어 결과가 의미없는 실험과, 실험의 주장에 대한 직교성

Main Claims of J&W (attention is not explanation)

- Attention weight는 다른 설명력 측정 수단인 feature importance measures(e.g. gradient-based measures)와 상관관계가 낮다.
  - Attention score가 trained model에 반영된다는 점에서 측정되는 상관관계의 level에서 의문이 생길 수 있지만, 이 주장에는 동의

- 하나의 예측에 대한 Attention weight의 구성은 여러 가지가 존재할 수 있으므로, 이는 설명력이 있다(plausible as explanation)고 볼 수 없다.
  - inferential perspective : ex) 오늘 비가 내린다면, 그 이유에 대한 설명에는 해류, 대기압, 구름 형성과 관련된 다양한 요소가 존재. 또한 '천둥의 신이 분노하였다' 라고도 설명 가능. 위의 설명들은 동일한 예측을 산출하지만 각 설명들이 동일한 설득력을 갖고 있진 않다.

Correlation is not Correlation

J&W's claim for second experiment

- 7개의 binary prediction datasets
- Attention score와 다른 interpretability measure들 간의 상관관계(gradient analysis, leave-one-out) 분석
- 동일한 예측을 내놓는 서로 다른 attention score 구성에 대한 적대적 탐색

Dataset	Avg. Length (tokens)	Train Size (neg/pos)	Test Size (neg/pos)
Diabetes	1858	6381/1353	1295/319
Anemia	2188	1847/3251	460/802
IMDb	179	12500/12500	2184/2172
SST	19	3034/3321	863/862
AgNews	36	30000/30000	1900/1900
20News	115	716/710	151/183

W&P's excuse(반박)

- 실험에서 correlation metric으로 설정한 Kendall-tau는 attention의 score value를 고려하지 않음.
  - Soft Attention distribution을 이용한 contextual model에 적합하지 않다.
  - Kendall-tau의 측정 방법을 고려했을 때, 7개의 dataset중 1개의 dataset(20News)만 0.33보다 낮음. 이는 여러 가지의 Attention score distribution이 상관관계 없이 좋은 예측을 할 수 있다고 볼 수 있다.
- gradient analysis, leave-one-out 외의 다른 평가지표
  - ex) human evaluation : Explainable Prediction of Medical Codes from Clinical Text(2018, Mullenbach et al.) attention score가 제일 informative하다.
  - J&W가 이와 같은 지표를 추가하여 설득력을 강화할 수 있었다.

Counterfactual Distributions are not Counterfactual Weights

Existence does not Entail Exclusivity

- Attention scores are claimed to provide an explanation; not the explanation.
  - 모델의 output layer의 같은 prediction에 대한 여러 attention weighting distribution 존재
  -> 모델의 동일한 예측에 대해 다른 여러 토큰에 가중치 부여에 기뻐해야 한다. (앞선 비 내리는 상황 인용)

Attention Distribution is not a Primitive.

- From a modeling perspective, detaching the attention scores obtained by parts of the model degrades the model itself.
  - model의 일부분에서 얻은 Attention score를 분리하는 것은 model의 품질을 저하시킨다.
- Attention weights where not assigned in some post-hoc manner by the model as the permutaion protocol assumes.
  - J&W는 Attention을 모델에 독립적인 존재로 간주하여 실험 결과를 도출함. 그러나, Attention은 모델의 구조에 종속되어 모델의 다른 layer와 연관하여 계산되는 구조이다.
  - J&W의 실험에서 모델에 대한 setup이 일관되지 않음. 즉, 기존에 train된 모델과 같은 예측을 내놓는 일관된 adversarial model에 대한 증거가 없다. -> 실험의 자유도 문제

Explanation is not Explanation

- Explanations is a loaded term, both in official AI terminology and in human conceptual thought.
  - Explanation의 정의에 대한 논의가 계속되고 있음.
  - 최근의 논문들에서는

- Most notable is Zachary Lipton's 2016 survey, The Mythos of Model Interpretability.
  - Explanation의 의미 : Interpretability(해석 가능성), Transparency(투명성)

- J&W cited it, but still use the terms somewhat interchangeably, making it hard to understand
  - 온전한 점검(sanity check)으로서의 Attention
    source text -> target text로의 mapping 과정에서 모델이 우리가 기대하는 패턴을 정확히 보여주는 것.
  - 도구(tool)로서의 Attention
    환자 방문 요약 : 모델이 상태 진단 예측. 모델의 출력은 의료 진단을 위한 도움이 될 수 있음.

Conclusion

- Attention might be explanation.
- It might not be explanation.

- Attentions == explanation의 여부는 Attention이 사용되는 모델의 아키텍처에 따라 결정된다.