



05_Explaining Language Models: Key Methods and Evaluations

1. Feature Attribution Methods

LLM에 대해서 설명 가능성이 왜 필요한가?
만약 언어 모델의 투명성이 부족하다면, 잘못된 정보나 유해한 콘텐츠를 생성할 가능성이 높아진다. 즉, LLM이 내부적으로 어떻게 작동하는지 명확히 이해되지 않은 상태에서 응답을 생성하기 때문에 신뢰할 수 없는 정보가 포함될 위험이 있는 것이다.
LLM의 설명 가능성을 높이는 것이 중요한 이유를 사용자에게 따른 구분을 통해 알아보자.

- 1. 일반 사용자(General Users)
 - 신뢰 확보
 - 모델의 능력과 한계를 파악 가능
- 2. 연구자 및 개발자(Researchers & Developers)
 - 모델의 편향과 성능 개선 영역 분석 가능
 - 디버깅 도구로 활용 가능

Feature Attribution

Feature Attribution은 말 그대로 기여도를 측정하는 것이다. ML/DL 모델의 예측 결과를 해석하는 과정에서 각 입력 특성(feature)이 예측에 얼마나 중요한 역할을 하는지 측정하는 기법이다.

- 입력 텍스트 x 가 n 개의 단어(feature) $\{x_1, x_2, ..., x_n\}$ 로 구성됨.
- 언어 모델 $f(x)$ 는 입력 x 에 대해 출력을 생성함.
- 특정 특성(단어) x_i 가 모델의 예측에 얼마나 중요한지 Relevance Score $R(x_i)$ 을 할당하여 측정.

| Method | 설명 |
|-------------------------------|--|
| Perturbation-based methods | 입력을 변경(perturbation)하면서 예측 변화 분석 (예 : Leave-one-out) |
| Surrogate Models | 원래 모델을 단순한 해석 가능 모델로 근사하여 설명 (예 : LIME, SHAP) |
| Backpropagation-based methods | 역전파를 활용하여 입력 특성이 예측에 미치는 영향 분석 |
| ➡ Gradient-based methods | 기울기를 기반으로 중요도를 계산 (예 : Saliency Maps) |


| | |
|-----------------------------|--|
| ➡ Propagation-based methods | 신경망을 통해 특성 중요도를 전파하여 측정 (예 : Layer-wise Relevance Propagation) |
| Attention-based methods | Transformer모델에서 Attention Score를 기반으로 중요한 특성 분석 |

Perturbation

1. Reduced Inputs(입력 축소 기법)

입력 데이터를 일부 제거해도 모델이 동일한 예측을 내리는지 확인하는 방법을 살펴보자. 사람이 이해하기 어려운 비논리적인 입력(reduced input)에서도 신경망 모델은 높은 신뢰도로 예측을 수행하는 문제를 탐색할 수 있다.

아래의 SNLI와 VQA 모델 사례를 보면, 사람이 보기에는 질문이 의미를 잃었지만, 모델은 여전히 높은 확신을 갖는 것을 볼 수 있다.

| | |
|---|--|
| SNLI | |
| Premise | Well dressed man and woman dancing in the street |
| Original | Two man is dancing on the street |
| Reduced | dancing |
| Answer | Contradiction |
| Confidence | 0.977 → 0.706 |
| VQA | |
|  | |
| Original | What color is the flower ? |
| Reduced | flower ? |
| Answer | yellow |
| Confidence | 0.827 → 0.819 |
| S. Feng(2018) Pathologies of Neural Models Make Interpretations Difficult | |

→ 즉, 모델이 논리적이지 않은 축소된 입력에서도 높은 신뢰도로 예측을 수행하면, 모델이 **불필요한 특징(irrelevant features)**에 의존하고 있을 가능성이 높은 것이다! 위의 과정을 통해서 모델의 설명가능성을 검증하고 편향을 찾아낼 수 있다!

2. Perturbation-based Explanation

Perturbation이란 특정 특성을 제거하거나 변형하여 모델의 예측 변화를 분석하는 기법이다. 특정 특성을 지우거나 바꿈으로써 모델의 예측이 어떻게 변하는지 살펴보면, 해당 특성이 얼마나 중요한지를 판단할 수 있다.

1. Leave-one-out 방식

- 하나의 특성을 제거하고 예측 결과 변화를 관찰
- Ex) 특정 단어나 토큰을 삭제하고 모델의 예측이 변하는지 확인

2. 특성 수준에서 변화 측정

- 단어, 토큰, 문장 스펠, 임베딩 벡터, 히든 유닛 등 다양한 수준에서 교란 수행

하지만, 이러한 Perturbation method에 문제점이 발생할 수 있다.

📌 Out-of-distribution(OOD) 문제

- 원래 데이터 분포에서 벗어난 비정상적인 입력이 생성될 수 있다.
- 문장에서 중요한 단어를 제거하면 의미가 완전히 달라질 수 있는 등의 예시.

📌 해결책

- 완전히 제거하는 대신, **작은 변화를 주는 방식**을 사용하여 원래 분포를 유지.

3. Perturbation-based Example

Perturbation-based

- Example

(input text) This movie was incredibly thrilling and well-acted!

(prediction) **positive**

(perturbation)

- This movie was incredibly [thrilling] and well-acted! → **positive**
- This movie was incredibly thrilling and [well-acted]! → **negative**

Surrogate Models

복잡한 블랙박스 모델의 예측을 설명하기 위해 더 단순하고 이해하기 쉬운 모델을 사용하는 방법이다. 대표적이 예시로 LIME, SHAP 등이 있다. 이런 모델들은 원래 모델의 개별 예측을 분석하여, 특정 입력 특성이 예측에 미치는 영향을 정량적으로 평가할 수 있다.

1. SHAP

- 특성 제거를 통해 해당 특성이 모델 예측에 미치는 영향을 평가하는 방법.
- 게임 이론의 Shapley Value를 기반으로 개별 특성의 기여도를 계산하는 기법.

📌 SHAP의 특징

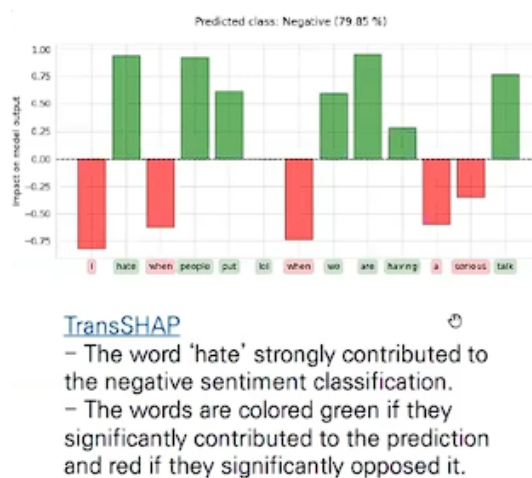
- 특성 제거
 - 입력 데이터에서 특정 특성을 제거할 때, 해당 값을 0, 평균(mean), 또는 다른 샘플 값으로 대체하여 제거 효과를 측정.
- 기준값(Baseline) 선택 문제
 - SHAP를 사용할 때 올바른 기준값 선택이 모호함.
 - Ex) 텍스트 데이터에서 특정 단어를 제거하면, 그 단어의 영향력을 어떻게 평가해야 할지 불분명함.
- 계산량이 많음
 - 특성 개수에 따라 계산량이 기하급수적으로 증가함.

2. TransSHAP (Transformer 기반 SHAP)

- SHAP를 자연어 처리 모델에 적용하는 방법.
- 단어 순서를 고려한 순차적 시각화(Sequential Visualization Explanation)를 제공.

📌 TransSHAP 특징

- 입력된 단어들이 모델 예측에 미치는 영향을 색상으로 시각화
 - Green : 해당 단어가 예측에 긍정적 영향을 미침
 - Red : 해당 단어가 예측에 부정적 영향을 미침
- 아래의 감정 분석 모델에서 TransSHAP를 사용한 예제를 보자.
 - 문장에서 "hate"라는 단어가 부정적 감정 예측에 강한 영향을 줌.
 - 즉, 이 단어가 없었다면 모델의 예측이 달라질 가능성이 있음.



Backpropagation-based

출력층에서 입력 특성까지 중요도(기여도) 점수를 전파하는 방식이다. 신경망이 학습할 때 사용하는 역전파 알고리즘을 활용하여 각 입력의 중요도를 계산한다.

1. Gradient Methods (기울기 기반 방법)

- 출력값 $f(x)$ 을 입력값 x 에 대한 편미분 계산.
- 기울기가 클수록 해당 특성이 모델의 예측에 더 큰 영향을 미친다고 판단.
- 입력 특성은 시각 모델에서는 pixel, NLP 모델에서는 token이 될 수 있다.

2. Propagation Methods (전파 기반 방법)

- 특정한 규칙을 정의하여 Layer-by-Layer로 기여도를 전파하는 방법.
- 일반적인 역전파와 다르게, 각 레이어마다 특화된 전파 규칙을 적용 가능.
- 출력에서 입력까지 단계적으로 특성 중요도를 계산.
- LRP, DeepLIFT 등이 대표적.

Vanila Gradients vs Integrated Gradients

1. Vanila Gradients

- 기울기의 부호 : 특정 특성이 예측에 긍정적/부정적으로 기여하는지 판단
- 기울기의 크기 : 해당 특성이 예측에 미치는 영향의 정도

📌 문제점

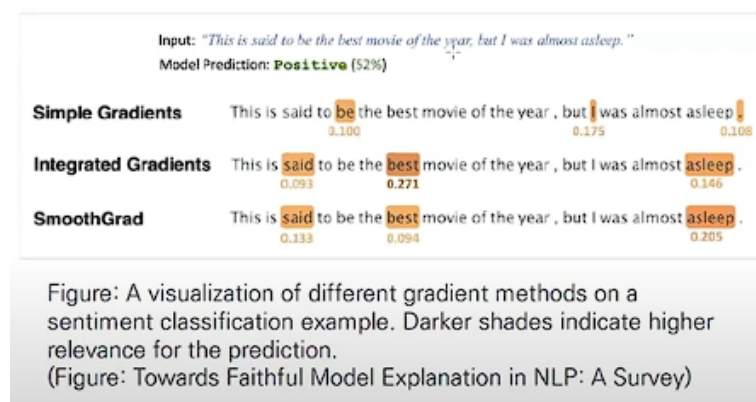
- 기울기는 특성이 예측에 얼마나 기여했는지(X)가 아니라, **출력의 민감도를 나타내는 것 뿐**이다.
- 단순 기울기 방식은 특성이 모델에 미치는 기여도를 올바르게 해석하지 못할 수 있다.
- Ex) 이미지 분류 (고양이 vs 강아지)
 - 신경망 모델이 이미지가 고양이인지 강아지인지를 예측하는 분류 모델.
 - 모델이 이미지의 각 픽셀(특성)에 대한 기울기를 계산하면,

"어떤 픽셀을 바꾸었을 때, 예측 값(고양이 확률)이 얼마나 변화하는지"를 측정할 수 있다.

- 즉, 기울기가 크다는 것은, 그 부분이 모델의 출력에 민감하게 반응하는 요소라는 것이지, 해당 특성이 반드시 "고양이"라고 분류한 이유는 아니다!!!

2. Integrated Gradients

- 입력 x 를 기준값과 비교하여 전체적인 기울기 변화를 적분하는 방법.
- 단순 기울기보다 특성이 예측에 미친 기여도를 더 정확하게 평가 가능.
- 기준값(Baseline)을 설정한 후, 입력을 기준값에서 점진적으로 변화시키며 기울기를 적분.
- NLP에서는 Zero Embeddings을 기준값으로 사용하여 단어의 기여도를 평가할 수 있음. 주로 문장 분류와 같은 순차적 데이터에 활용.



Layer-wise Relevance Propagation (LRP)

신경망의 각 레이어별로 출력에서 입력까지 중요도를 역전파하는 방식.

Relevance Conservation Constraint을 도입하여

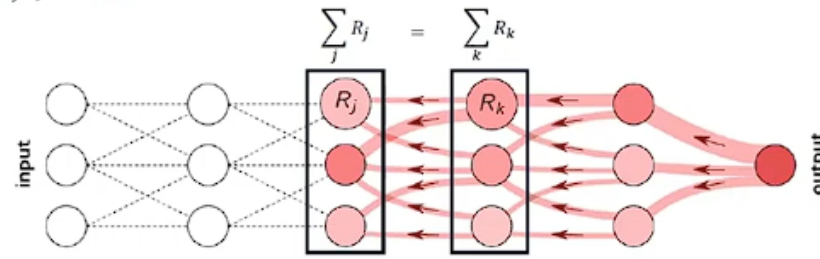
각 뉴런에 들어오는 총 기여도 = 각 뉴런에서 나가는 총 기여도를 유지한다.

- 네트워크의 각 뉴런이 모델의 최종 예측에 기여하는 방식을 추적할 수 있다.
- 특정 뉴런이 최종 예측에 얼마나 중요한지를 평가 가능.

Conservation property

- The sum of the relevance scores remains the same across all layers.

$$\sum_j R_j = \sum_k R_k$$



Explaining sentiment analysis - IG, LRP

(input) I found it an almost perfect film, with some deliciously carefully crafted moments and great acting.

[CLS] i found it an almost perfect film, with some deliciously carefully crafted moments and great acting. [SEP]

Model: Bert
Label / Prediction: Positive / Positive
Confidence (softmax): 0.9996
Explainer: IntegratedGradients

[CLS] i found it an almost perfect film, with some deliciously carefully crafted moments and great acting. [SEP]

Model: Bert
Label / Prediction: Positive / Positive
Confidence (softmax): 0.9996
Explainer: LRPUniformEpsilon

Which explainer provides a better explanation (more faithful explanation)?

Explaining sentiment analysis - LIME, KernelSHAP

(input) I found it an almost perfect film, with some deliciously carefully crafted moments and great acting.

[CLS] i found it an almost perfect film, with some deliciously carefully crafted moments and great acting. [SEP]

Model: Bert
Label / Prediction: Positive / Positive
Confidence (softmax): 0.9996
Explainer: Lime

[CLS] i found it an almost perfect film, with some deliciously carefully crafted moments and great acting. [SEP]

Model: Bert
Label / Prediction: Positive / Positive
Confidence (softmax): 0.9996
Explainer: KernelShap

Which explainer provides a better explanation (more faithful explanation)?

Propagation Methods - Strengths & Weakness

장점 (Strengths)

- 다양한 **Feature Relevance Score**(특성 중요도 점수)를 생성 가능.
- 일부 방법(예: LRP, DeepLIFT)은 기존 기법보다 더 높은 신뢰도를 제공.
- 계산량이 다양하지만, 일부 방법은 상대적으로 빠르게 수행될 수 있음.

단점 (Weaknesses)

- 대부분의 역전파 기반 기법은 낮은 수준의 특성(픽셀, 토큰)에 초점.
- 고차원적인 특성(예: 성별, 문법적 의존 관계)에 대한 기율기 계산이 어려움.
- 텍스트 생성 모델(Text Generation) 등에는 적용하기 어려움.
- 일부 방법은 설명이 **불안정(unstable)**하여, 입력이 조금만 변해도 설명이 크게 달라질 수 있음.
- 일부 기법은 설명의 신뢰성(Faithfulness)을 완벽히 보장하지 못할 수도 있음.

2. Analysis of model-internal structure

Analysis on Neurons

신경망 내부에서 **특정 뉴런이 특정한 역할을 수행하는 패턴**을 분석한 연구.

아래 그림을 보면, 뉴런들이 '따옴표 내부'에서 활성화되는 경향을 보인다.

이와 같이 '따옴표 내부', 'if문 내부', '문장의 끝부분'에서 활성화되는 경향을 보이고, 해당 뉴런이 어떤 의미를 학습했는지 이해하는 것이 모델 해석에 중요한 역할을 한다.

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chicago, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.
Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

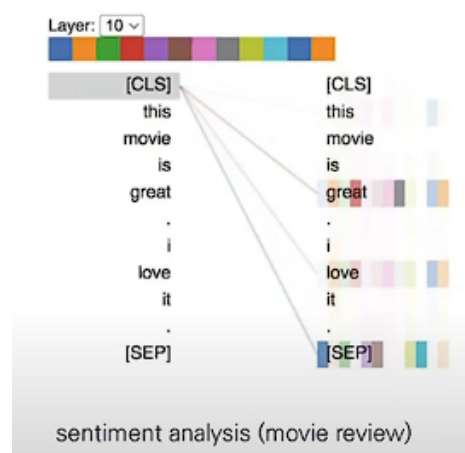
Figure 2
A neuron that "turns on" inside quotes (figure from Karpathy, Johnson, and Fei-Fei 2015). Blue/red indicates positive/negative activations, respectively, and a darker shade indicates larger magnitude.

From a character-level LSTM language model. Karpathy, Johnson, and Fei-Fei (2015)

Analysis on Attention Mechanism

아래의 BERT 모델을 이용한 감성 분석 예제를 살펴보자.

- 각 색깔은 Attention Head를 나타내며, [CLS] 토큰이 모든 토큰과의 어텐션 가중치를 보여준다.
- 더 어두운 색상일수록 더 높은 어텐션 가중치를 의미하며, 'great'와 'love'가 높은 가중치를 받은 것을 확인할 수 있다.
- 이러한 어텐션 메커니즘을 통해 모델이 어떤 단어를 중요하게 여기는지 확인할 수 있다.



Attention is not Explanation

하지만, 어텐션이 반드시 모델의 설명력을 보장하지 않는다는 주장이 거론되었다.

아래의 같은 문장에 대해, 어텐션 가중치를 다르게 조작(adversarial attention) 했음에도 불구하고, 모델의 예측이 동일하다.

즉, 모델이 특정 단어에 높은 어텐션을 할당한다고 해서, 그것이 반드시 예측에 중요한 요소라고 단정할 수 없기 때문이다.

| | |
|---|--|
| after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore | after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore |
| original α | adversarial $\tilde{\alpha}$ |
| $f(x \alpha, \theta) = 0.01$ | $f(x \tilde{\alpha}, \theta) = 0.01$ |

Debate over Attention

어텐션이 설명력이 있는가? 에 대한 논쟁이 있었다.

- Jain & Wallace (2019) : “Attention is Not Explanation”
 - 조작된 어텐션 가중치로도 같은 결과가 나오기 때문에, 어텐션이 모델이 예측하는 근거라고 보기 어렵다.
- Wiegrefe (2019) : “어텐션이 항상 설명이 불가능한 것은 아니다”
 - 특정한 상황에서는 신뢰할 수 있는 설명이 될 수 있음.

[Reconciliation]

두 연구(논쟁 당사자들) 모두, “LSTM 네트워크에서 어텐션이 특정 조건에서는 신뢰할 수 있는 설명 역할을 할 수 있다”고 인정한다.

✚ 하지만 일반적인 법칙(one-size-fits-all)으로 적용하기는 어렵다.

✚ 어텐션이 개별 샘플(instance)에 대한 설명으로는 신뢰도가 떨어진다. 반드시 예측 결과의 원인을 설명하는 것이 아니기 때문이다.

✚ 그럼에도 Transformer 구조의 전반적인 동작 방식을 분석하는 데는 여전히 중요한 역할을 한다.

Why is Attention Unfaithful?

어텐션이 설명력(Faithfulness)을 갖지 않는 이유:

1. 정보 혼합(Information Mixing)

- 어텐션 가중치는 입력이 아니라 은닉 상태(hidden states)에 할당됨.
- 초기 계층에서 이미 정보가 혼합되어, 나중의 어텐션 값이 원본 정보와 직접적인 관계가 없음.

2. 지역성(Locality)

- 대부분의 어텐션 연구는 단일 계층의 어텐션 가중치만 분석.
- 이는 전체 네트워크에서 정보가 어떻게 흐르는지를 반영하지 못하는 단점이 있음.

3. 인과 관계 부족(Intrinsic Lack of Causality)

- 어텐션 가중치는 예측에 대한 직접적인 인과 관계를 반영하지 않음.
- 다른 설명 방법(ex: 역전파 기반 방법)과 함께 사용해야 신뢰도를 높일 수 있음.

Approaches to Tackle the Issues

어텐션의 문제점을 해결하는 방법:

1. 정보 혼합 문제 해결

- 은닉 상태와 입력 특징 간의 거리를 최소화(Weight Tying)
- 은닉 상태가 원래 입력을 더 잘 대표하도록 설계.

2. 지역성(Locality) 문제 해결

- 전체 네트워크에서 정보가 어떻게 흐르는지 반영하는 방법 개발.
- Attention Rollout 및 Attention Flow 기법(Abnar & Zuidema, 2020) 도입.
- 기존 어텐션 가중치보다 신뢰성이 높다고 입증됨.

3. 인과 관계 부족 해결

- 어텐션을 다른 설명 방법과 결합하여 사용.

Attention Rollout

Attention Rollout은 Transformer 모델에서 **토큰 간 정보 흐름을 보다 신뢰성 있게 추적**하기 위한 방법이다.

일반적인 self-attention layer에서는 정보가 여러 계층에서 점점 섞이며 최종적으로 해석하기 어려운 attention 맵이 생성된다. 특히,

3개 이상의 계층을 지나면 attention 맵이 의미가 없는 상태가 되기 때문에, 이를 개선하기 위해 Attention Rollout이 제안되었다.

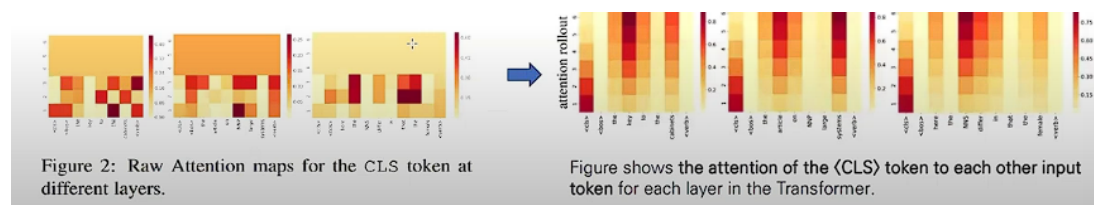
[핵심 원리]

- Attention Rollout은 각 층의 attention을 누적(dot product)하여 최종 attention을 계산한다.
- Residual connection을 고려하여 토큰 간의 영향을 더 정확하게 측정한다.
- 이를 통해 최종적으로 각 입력 토큰과 특정 토큰(ex: [CLS]) 간의 중요도를 보다 명확하게 파악할 수 있다.

$$\bar{A}(l_i) = \begin{cases} A(l_i) \bar{A}(l_{i-1}) & \text{if } i > j \\ A(l_i) & \text{if } i = j \end{cases}$$

attention rollout of current layer
 attention matrix of current layer
 attention rollout of previous layer

- 위 수식에서:
 - $A(l_i)$: 현재 layer의 attention 행렬
 - $\bar{A}(l_{i-1})$: 이전 layer의 attention rollout 값
 - 현재 층의 attention 행렬을 이전 층의 attention과 연속적으로 곱하여, 전체 네트워크에서 정보가 어떻게 흐르는지 추적한다.

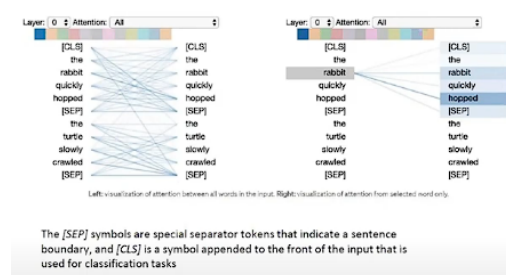


Attention Visualization

- Transformer 기반 모델(BERT, GPT-2, T5 등)의 Attention을 시각화하는 방법을 보자.
- Attention Visualization을 통해 모델이 특정 예측을 수행할 때 어떤 단어에 집중하는지 확인할 수 있다.
- 이를 활용하면 모델의 편향(Bias), 오류(Error), 의사결정 규칙(Decision Rules) 등을 분석하는데 도움이 된다.

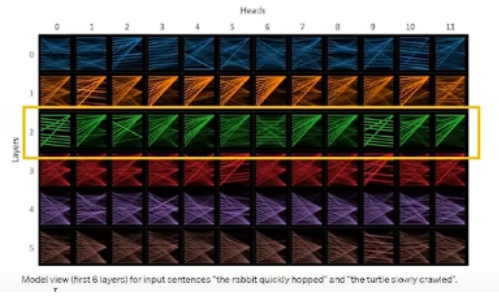
[Attention Map 예제]

- 아래 이미지는 Attention이 어떻게 작동하는지를 시각적으로 보여준다.
 - 좌측 : 입력 문장의 모든 단어들 간의 Attention 연결을 보여줌.
 - 우측 : 선택한 특정 단어에 대한 Attention을 보여줌.
- [CLS] : 문장 분류에서 사용되는 특별한 시작 토큰
- [SEP] : 문장의 경계를 나타내는 특별한 토큰



[Multi-head Attention Visualization]

- 아래 이미지는 여러 Attention Head(12개)의 패턴을 시각적으로 보여준다.
- 각 레이어와 Head에서 특정 단어들이 서로 어떻게 Attention을 주고받는지 분석 가능.
- 노란색 박스 : 모델이 “the rabbit quickly hopped”와 “the turtle slowly crawled” 문장에 대해 “Next-word Attention” 패턴을 보이는 것을 보여줌.
 - 즉, 현재 단어가 아니라 다음 단어로 Attention이 집중된다.
 - Transformer가 문맥을 이해할 때 다음 단어가 현재 단어의 의미를 결정하는 데 중요한 역할을 한다는 것을 시사한다.



Analysis of model-internal structures Strengths & Weaknesses

1. Strenghts

- 직관적이고 사람이 이해하기 쉬운 시각화 가능
→ 모델 내부의 구조를 시각적으로 표현하면 직관적이고 해석이 용이함
- 인터랙티브 툴 지원(e.g. BertViz)
→ 사용자가 데이터를 탐색하고 가설을 세우는 데 도움을 주는 다양한 툴이 존재함
- 어텐션 메커니즘이 특징 간 상호작용을 포착할 수 있음
→ 대부분의 기법들은 개별 특징의 중요도만을 측정하는 반면, 어텐션은 여러 특징 간의 관계를 반영할 수 있음
- 모델 가중치에 쉽게 접근할 수 있으며 계산적으로 효율적
→ 모델이 학습한 가중치를 직접 확인할 수 있고, 계산 비용도 상대적으로 낮음

2. Weaknesses

- 어텐션 가중치가 인과적(contributive) 영향을 얼마나 반영하는지는 불분명
→ 단순히 어텐션이 높다고 해서 모델의 최종 예측에 직접적인 기여를 했다고 단정할 수 없음
- 신뢰성(Faithfulness) 부족 문제
→ 어텐션이 입력 특징의 중요도를 측정하는 것이 아니라 중간(hidden states)에서의 어텐션을 해석하기 때문에 오해의 소지가 있음
- 대부분의 기법들이 단일 레이어 또는 단일 토큰 위치에서의 어텐션 가중치만 고려함
→ 이는 모델이 특정 입력 위치에 얼마나 집중했는지를 반영할 수 있지만, 전체적인 연산 경로를 고려하지 않음

3. Evaluation of Explanations

설명에 대한 평가에 대해서 알아보자.

1. 사용자 연구 (User Study)

- 최종 사용자가 여러 설명 방법을 비교.
- 비용이 많이 들고 시간이 오래 걸림.
- 인간의 평가에는 편향이 개입될 가능성이 크다.

2. 정답 기반 평가 (Ground Truth-Based Evaluation)

- 인간 주석자(annotators)가 제공한 정답 해석을 근거라고 부름.
- 텍스트 분류에서 이러한 근거는 단어, 문장 또는 텍스트의 특정 부분이 될 수 있음.
- 근거가 존재하는 경우, 표준적인 성능 지표(F1, AUPRC)를 활용하여 해석을 평가할 수 있음.
- 그러나, 이러한 근거가 포함된 데이터셋은 매우 드물고, 경우에 따라 오류가 있거나 편향될 수 있음.

Faithfulness

신뢰성은 설명의 가장 근본적인 요구사항이다. 설명은 모델의 예측 과정을 정확히 반영해야하고, 가장 중요한 특성을 제거했을 때 예측이 얼마나 바뀌는지를 측정했을 때 변화가 클수록 더 신뢰할 수 있는 설명이라고 할 수 있다.

Faithfulness Metrics

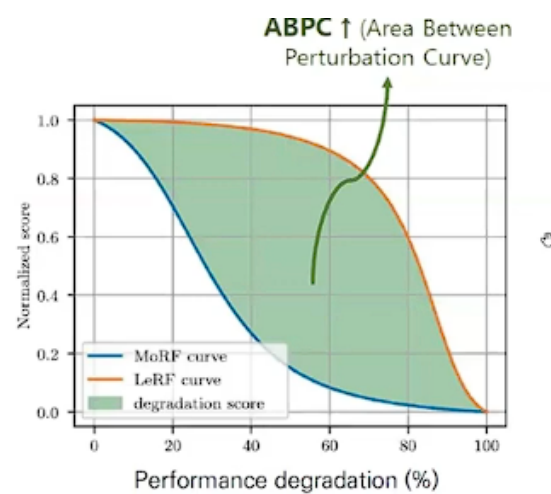
아래 그림과 같은 신뢰성 평가 지표에 대해 알아보자.

1. AOPC (Area Over Perturbation Curve)
 - 중요한 특징을 제거하면서 예측 변화량을 곡선 아래 면적으로 계산.
2. ABPC (Area Between Perturbation Curve)
 - 중요한 특징을 제거할 때 성능이 어떻게 저하되는지 평가.

[그래프 해석]

- MoRF (Most Relevant First) : 가장 중요한 특징부터 제거.
- LeRF (Least Relevant First) : 가장 덜 중요한 특징부터 제거.

→ **MoRF 곡선이 빠르게 하락할수록 설명의 신뢰성이 높다.**



Comprehensiveness & Sufficiency

- 포괄성 : 설명에서 제공한 근거(Rationale)를 제거하면 모델의 신뢰도가 감소해야 함.
- 충분성 : 모델이 해당 근거만을 이용해서 원래 예측과 동일한 결론을 내릴 수 있어야 함.

결론적으로, 신뢰성이 높은 설명일수록 핵심적인 특성을 제거하면 예측 성능이 큰 영향을 받아야 한다! 또한, 근거만으로도 올바른 예측을 할 수 있어야 한다!!

Research Challenges

1. 신뢰성 평가의 어려움

- 신뢰성을 객관적으로 평가하는 명확한 기준이 없음.
- 보편적인 평가 프레임워크의 필요성.

2. 설명 기법의 한계

- 대부분의 방법이 개별 특성(feature)의 기여도를 측정할 뿐, 고차원적인 특성 간의 상호작용을 포착하지 못함.
→ 예) 문법, 의미론, 담화 구조 등 고려 부족.

3. 연구 초점의 편향

- 기존 연구는 주로 시퀀스 분류(sequence classification)에 집중됨.
→ 생성 모델(Generative Language Models)에 대한 연구 필요.

4. 신뢰성(Faithfulness)만이 목표가 아님

- 설명이 단순히 신뢰성만 높은 것이 아니라, **사용자의 의사결정, 모델 디버깅, 지식 발견 등 실질적인 도움을 줄 수 있어야 함.**

📌 **최종 목표:**

- **투명성(Transparency), 해석 가능성(Explainability), 안전성(Safety)을 모두 고려하여 신뢰할 수 있는 LLM(대형 언어 모델) 구축이 필요!**